ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

«Ανάπτυξη βιοπληροφορικών ροών για την μεταγονιδιωματική ανάλυση κλινικών δεδομένων και ενσωμάτωση τους σε διαδραστική πλατφόρμα»

Στυλιανή Αναγνώστου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος Καθηγητής:
Χαράλαμπος Καρανίκας
Επίκουρος Καθηγητής

Επιστημονικός Υπεύθυνος Ε.Ι.Παστέρ:
Τιμοκράτης Καραμήτρος
Εντεταλμένος Ερευνητής

Λαμία, Ιούνιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

«Ανάπτυξη βιοπληροφορικών ροών για την μεταγονιδιωματική ανάλυση κλινικών δεδομένων και ενσωμάτωση τους σε διαδραστική πλατφόρμα»

Στυλιανή Αναγνώστου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων Καθηγητής:
Χαράλαμπος Καρανίκας
Επίκουρος Καθηγητής

Επιστημονικός Υπεύθυνος Ε.Ι.Παστέρ:
Τιμοκράτης Καραμήτρος
Εντεταλμένος Ερευνητής

Λαμία, Ιούνιος 2022

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

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

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

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

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

Ημερομηνία: 15/06/2022

Ο – Η Δηλ.

(Υπογραφή)

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

Ανάπτυξη βιοπληροφορικών ροών για την μεταγονιδιωματική ανάλυση κλινικών δεδομένων και ενσωμάτωση τους σε διαδραστική πλατφόρμα

Στυλιανή Αναγνώστου

Τριμελής Επιτροπή:

Καρανίκας Χαράλαμπος, Επίκουρος Καθηγητής (επιβλέπων)

Κακαρούντας Αθανάσιος, Αναπληρωτής Καθηγητής

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Ευχαριστίες

Η παρούσα πτυχιακή εργασία εκπονήθηκε κατά το χρονικό διάστημα 2021-2022, στη Μονάδα Βιοπληροφορικής και Εφαρμοσμένης Γενωμικής του Ελληνικού Ινστιτούτου Παστέρ σε συνεργασία με το τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας.

Αρχικά, θα ήθελα να ευχαριστήσω για τη συμμετοχή τους τα μέλη της τριμελούς εξεταστικής επιτροπής, τον Αναπληρωτή Καθηγητή Αθανάσιο Κακαρούντα, τον Επίκουρο Καθηγητή Σωτήριο Τασουλή και ιδιαίτερα τον Επίκουρο Καθηγητή Χαράλαμπο Καρανίκα για την ευκαιρία συνεργασίας στο πλαίσιο της πτυχιακής μου εργασίας και για την υποστήριξη και συμβουλές του.

Οφείλω ένα μεγάλο ευχαριστώ στον επιστημονικό υπεύθυνο της παρούσας πτυχιακής εργασίας, τον Εντεταλμένο Ερευνητή της Μονάδας Βιοπληροφορικής και Εφαρμοσμένης Γενωμικής, του Ε.Ι.Π., Δρ. Τιμοκράτη Καραμήτρο, ο οποίος με δέχτηκε στην ερευνητική του ομάδα και με καθοδήγησε από την αρχή μέχρι την ολοκλήρωση του εγχειρήματος. Εκτιμώ βαθιά την εμπιστοσύνη, τις συμβουλές και την ευκαιρία να εργαστώ μαζί του. Θα ήθελα ακόμα να ευχαριστήσω το προσωπικό του εργαστηρίου για το ευχάριστο κλίμα εργασίας, τις συμβουλές και την ψυχολογική στήριξη που μου προσέφεραν στον καιρό συνεργασίας μας, τις υποψήφιες διδάκτορες Μαρία Μπούσαλη και Γεθσημανή Παπαδοπούλου και τους μεταπτυχιακούς φοιτητές Αριστέα Δημάδη και Δημήτριο-Χρήστο Τρεμούλη.

Τέλος, ένα τεράστιο ευχαριστώ οφείλω στην οικογένεια και τους φίλους μου για την συνεχόμενη στήριξη και υπομονή τους.

Πίνακας Περιεχομένων

Περίληψη 6
Abstract 7
1. Εισαγωγή 8
1.1 Μεταγονιδιωματική 8
1.2 Κλινική Μεταγονιδιωματική 9
1.3 Αλληλούχιση Επόμενης Γενεάς - Next Generation Sequencing (NGS) 9
1.4 Προσεγγίσεις NGS στην μεταγονιδιωματική 10
1.5 Σύγκριση Μεταγονιδιωματικών Εργαλείων 13
1.6 Σκοπός 16
2. Μεθοδολογία 17
2.1 Προγραμματιστικά Εργαλεία 17
2.1.1 Python 17
2.1.2 Django 17
2.1.3 Galaxy 17
2.1.4 BioBlend 19
2.2.1 Kraken 2 21
2.2.2 Krona 22
2.3 Προϋπάρχουσες Εφαρμογές 24
2.3.1 NGPhylogeny.fr 24
2.4 Σύνολο δεδομένων Μεταγονιδιωματικής 24
3. Υλοποίηση Εφαρμογής 25
3.1 Κώδικας 25
3.1.1 Front-End 25
3.1.2 Back-End 25
3.2 Πρακτική Εφαρμογή 26
3.3 Σύγκριση των αποτελεσμάτων που λήφθηκαν από την χρήση της πλατφόρμας με τα αποτελέσματα που λήφθηκαν με χρήση του Terminal 30
4. Συμπεράσματα 34
Βιβλιογραφία 35

Περίληψη

Η μεταγονιδιωματική είναι μια σύγχρονη προσέγγιση στην διερεύνηση της σύνθεσης και της δυναμικής των μικροβιακών κοινοτήτων, που αναπτύσσεται διαρκώς τα τελευταία χρόνια χάριν στην πρόοδο που έχει επιτευχθεί στις τεχνολογίες αλληλούχισης επόμενης γενεάς (Next Generation Sequencing - NGS). Τα μεταγονιδιωματικά δεδομένα χαρακτηρίζονται ως δεδομένα «μεγάλου όγκου» (Big Data) και για την επεξεργασία τους απαιτούνται εξειδικευμένοι βιοπληροφορικοί αλγόριθμοι.

Σκοπός της παρούσας πτυχιακής εργασίας είναι η ανάπτυξη βιοπληροφορικών ροών μεταγονιδιωματικών αναλύσεων για την ανάλυση κλινικών δεδομένων μικροβιώματος και η ενσωμάτωση αυτών σε μια ελεύθερα προσβάσιμη διαδραστική πλατφόρμα.

Για την ανάπτυξη των βιοπληροφορικών ροών μεταγονιδιωματικής επιλέχθηκε το υπολογιστικό εργαλείο Kraken 2, ενώ για την ανάπτυξη της διαδραστικής πλατφόρμας χρησιμοποιήθηκε η δομή Galaxy και οι γλώσσες προγραμματισμού Python, HTML, CSS και JavaScript.

Μέσω της διαδραστικής πλατφόρμας και των μεταγονιδιωματικών ροών που αυτή εγκολπώνει, ο χρήστης έχει τη δυνατότητα ανάλυσης μεγάλου όγκου δεδομένων μέσω εξειδικευμένων ροών, χωρίς να απαιτείται πρότερη γνώση προγραμματισμού, ενώ παράλληλα έχει προβλεφθεί η δυνατότητα ενσωμάτωσης επιπλέον εργαλείων και ροών, ως μελλοντικός στόχος.

Abstract

Metagenomics is a modern approach in investigating the synthesis and dynamic of microbial communities, which has been developing for the past few years aided by the progress that has been made in Next Generation Sequencing (NGS) technologies. Metagenomic data are characterized as «Big Data» and advanced bioinformatic algorithms are required for their analysis.

The goal of this thesis is the development of bioinformatic pipelines for metagenomic analysis for the analysis of clinical microbiome data and their integration into a freely accessible interactive platform.

For the development of the metagenomics bioinformatic workflows the computational tool Kraken 2 was chosen, while for the development of the interactive platform we used Galaxy and Python, HTML, CSS and Javascript as programming languages.

Through the interactive platform and the metagenomic workflows that it embeds, the user has the ability to analyze large volumes of data through specialized pipelines, without the need for prior programming knowledge, while the possibility of integrating additional custom tools and workflows, is a future goal.

1.Εισαγωγή

1.1 Μεταγονιδιωματική

Πριν την ανάπτυξη των τεχνολογιών αλληλούχισης επόμενης γενεάς (Next Generation Sequencing - NGS), ο εντοπισμός και η ταξινόμηση των μικροοργανισμών απαιτούσε την πρότερη καλλιέργεια αυτών. Η εφαρμογή της μεταγονιδιωματικής έδωσε νέες προοπτικές στη μικροβιολογία διερευνώντας άμεσα τη σύνθεση μικροβιακής κοινότητας, επιτρέποντας την ταχύτερη ανίχνευση και ανακάλυψη νέων ειδών και μειώνοντας την εξάρτηση από προσεγγίσεις που εξαρτώνται από την καλλιέργεια. Παράλληλα, δόθηκε η δυνατότητα διερεύνησης της σύνθεσης και της δυναμικής των μικροβιακών κοινοτήτων που ανευρίσκονται σε ποικίλα οικοσυστήματα (χερσαία, υδάτινα, ζωικοί ιστοί κλπ)

Η μεταγονιδιωματική αποτελεί έναν σύγχρονο κλάδο της μοριακής βιολογίας, ο οποίος έχει ποικίλες εφαρμογές στη διερεύνηση των μικροβιακών κοινοτήτων που έχουν δειγματοληφθεί άμεσα από το περιβάλλον ή από κλινικά δείγματα χωρίς την πρότερη καλλιέργεια ή απομόνωση καθένος από τους μικροοργανισμούς που υπάρχουν στην υπό μελέτη κοινότητα. Ο λόγος προτίμησής της είναι το γεγονός ότι ορισμένα βακτήρια είναι δύσκολα στη καλλιέργεια ή μη-καλλιεργήσιμα. Επιπλέον, χρειάζονται μεγάλο χρόνο ανάπτυξης, ιδιαίτερα σε ασθενείς με λοιμώξεις χαμηλού βαθμού. Μέσω της μεταγονιδιωματικής αναγνωρίζονται και ταξινομούνται όλα τα μικροβιακά στελέχη που είναι παρόντα στην υπό μελέτη κοινότητα (περιβαλλοντικό ή κλινικό δείγμα) σε μεγαλύτερο βάθος και επιπλέον προσφέρονται λεπτομέρειες για τις μεταβολικές διεργασίες και τους λειτουργικούς ρόλους των μικροβίων στα δείγματα. α τελευταία χρόνια έχει αναπτυχθεί πληθώρα προγραμμάτων με γενικά καλή επίδοση δίνοντας έτσι τη δυνατότητα στους χρήστες την επιλογή να διαλέξουν το εργαλείο που ανταποκρίνεται περισσότερο στις ανάγκες και δυνατότητές τους (Ye et al., 2019). Ως αποτέλεσμα, η μεταγονιδιωματική ανάλυση έχει γίνει πιο προσεγγίσιμη από ποτέ.

Παρόλο που οι μεταγονιδιωματικές μέθοδοι δείχνουν ότι είναι πολλά υποσχόμενες στην ανάπτυξη στους τομείς της βιολογίας και της βιοπληροφορικής, υπάρχουν ακόμα κάποια όρια στις ικανότητες τους. Αρχικά δεν έχουν τη δυνατότητα διαχωρισμού μεταξύ ζώντων μικροοργανισμών και «αποτυπωμάτων» που μαρτυρούν παρελθούσα παρουσία αυτών και έχουν χαμηλή ικανότητα να ανιχνεύσουν μικροοργανισμούς με χαμηλή αναπαράσταση DNA στο δείγμα (Loeffler et al., 2019). Επιπλέον, εάν το δείγμα προέρχεται από άνθρωπο η μελέτη του μικροβιώματος καθίσταται πιο δύσκολη, διότι συνήθως αυτό το ανθρώπινο γονιδίωμα κυριαρχεί έναντι των γονιδιωμάτων των μικροοργανισμών.

1.2 Κλινική Μεταγονιδιωματική

Η μεταγονιδιωματική έχει εφαρμογές με σκοπό την βελτίωση της διαγνωστικής διεργασίας σε επίπεδο δημόσιας υγείας και ιατρικής βασισμένης σε τεκμήρια (evidence-based medicine), οι οποίες εντάσσονται στον τομέα της κλινικής μεταγονιδιωματικής. Η κλινική μεταγονιδιωματική αντλεί τις ρίζες της στην ευρύτερη χρήση μικροσυστοιχιών στις αρχές του 21ου αιώνα (Chiu and Miller, 2019) και είναι η διαδικασία αλληλούχισης νουκλεϊκών οξέων κλινικών δειγμάτων με σκοπό την άντληση κλινικά σημαντικής πληροφορίας όπως η ταυτοποίηση μικροοργανισμών και η ευαισθησία τους σε αντιμικροβιακά (d'Humières et al., 2021). Αυτό που βοήθησε περισσότερο στην ανάπτυξή της υπήρξε η ανάπτυξη του NGS. Ο κλάδος περιλαμβάνει επίσης τη διαγνωστική μικροβιολογία και τη μικροβιολογία δημόσιας υγείας.

Ενώ οι περισσότερες μοριακές αναλύσεις στοχεύουν μόνο σε έναν περιορισμένο αριθμό παθογόνων με χρήση ειδικών ανιχνευτών, οι μεταγονιδιωματικές προσεγγίσεις χαρακτηρίζουν όλο το DNA ή το RNA που υπάρχει σε ένα δείγμα, επιτρέποντας την ανάλυση ολόκληρου του μικροβιώματος καθώς και του γονιδιώματος ή του μεταγραφώματος του ξενιστή. Βασικό πλεονέκτημα της κλινικής μεταγονιδιωματικής σε σχέση με τις συμβατικές μεθόδους είναι η ικανότητα ανίχνευσης μικροοργανισμών με εξαντλητικό τρόπο χωρίς καμία προϋπόθεση, ο υπολογισμός των σχετικές αναλογίες τους και η παραγωγή πληροφοριών σχετικά με την ευαισθησία τους σε αντιμικροβιακά (d'Humières et al., 2021).

Εφαρμογές της κλινικής μεταγονιδιωματικής αποτελούν η ανίχνευση λοιμωδών νοσημάτων, χαρακτηρισμός του μικροβιώματος σε παθολογικές καταστάσεις έναντι των φυσιολογικών, χαρακτηρισμός της απόκρισης του ξενιστή σε μόλυνση. Ανάλυση δεδομένων μικροβιώματος εντέρου αναδεικνύουν ότι ποικίλες κοινότητες μικροοργανισμών μπορούν να επηρεάσουν την αντίδραση του ατόμου σε φαρμακευτική αγωγή καθώς και τη διάγνωση μολύνσεων κοκάλων και αρθρώσεων (National Research Council, 2007).

Δυσκολίες που παρουσιάζει η κλινική μεταγονιδιωματική σχετίζονται κυρίως με την ποσότητα του ανθρώπινου DNA που υπάρχει στα δείγματα σε σύγκριση με μικροβιακό DNA καθώς και το κόστος που απαιτείται για την πραγματοποίηση του NGS.

1.3 Αλληλούχιση Επόμενης Γενεάς - Next Generation Sequencing (NGS)

Οι τεχνολογίες Αλληλούχισης Επόμενης Γενεάς - Next Generation Sequencing (NGS) είναι μια μέθοδος υψηλής απόδοσης που επιτρέπει την εξαγωγή μεγάλου όγκου δεδομένων για πολλαπλά δείγματα παράλληλα κατά τη διάρκεια ενός πειράματος (run) (Levy and Myers, 2016). Η βασική μονάδα του NGS είναι η 'ανάγνωση' ('read'), αλληλουχία μικρής έκτασης, με σύνηθες μήκος 100-500 ζεύγη βάσεων (ζβ) Οι πιο διαδεδομένες NGS πλατφόρμες αυτη της Illumina, της ThermoFisher, της Nanopore, της BGI και της PacBio. Οι τελευταίες δύο παράγουν «μακριές» αναγνώσεις (long-reads) (<1000ζβ) και οι άλλες τρεις «κοντές» αναγνώσεις (short-reads) (>300ζβ). Η Illumina έχει την πρωτιά στην αγορά στις τεχνολογίες αλληλούχισης 'κοντών'

αναγνώσεων με το μικρότερο κόστος ανά ζεύγος γιγαβάσεων. Η PacBio και η Nanopore έχουν αναπτύξει NGS πλατφόρμες που παράγουν «μακριές» αναγνώσεις αλλά το υψηλό κόστος χρήσης και το μεγαλύτερο ποσοστό λαθών που τις χαρακτηρίζουν δεν τις καθιστούν καλύτερη επιλογή από την πλατφόρμα της Illumina (Loeffler et al., 2019).

Η αλληλούχιση επόμενης γενεάς αποτέλεσε αφετηρία για την ανάπτυξη της μεταγονιδιωματικής, επιτρέποντας της ανάλυση μεγάλου όγκου γονιδιωματικών δεδομένων προερχόμενων από κλινικά ή περιβαλλοντικά δείγματα. Η ανάπτυξη καινοτόμων και εξειδικευμένων εργαλείων για την ανάλυση των NGS δεδομενων καθώς και η κατακόρυφη μείωση του κόστους των ΝGS πειραμάτων, οδήγησαν στην γρήγορη υιοθέτηση της τεχνολογίας NGS σε μια σειρά πεδίων, συμπεριλαμβανομένης και της κλινικής διάγνωσης.

Η μεταγονιδιωματική αλληλούχιση επόμενης γενεάς (mNGS) χρησιμοποιείται σε κλινικό επίπεδο, περιλαμβάνοντας τον χαρακτηριστιρισμό αντιβιοτικής αντίστασης απευθείας από κλινικά δείγματα και ανάλυση των δεδομένων ανταπόκρισης του ανθρώπινου ξενιστή με σκοπό την πρόβλεψη των λόγων μόλυνσης και αξιολόγηση του κινδύνου ασθένειας. Έτσι, η mNGS μπορεί να υπάρξει βασικός φορέας για την ακριβής διάγνωση μεταδοτικών ασθενειών.

Όσον αφορά τις μεθόδους της τεχνολογίας χρησιμοποιούνται κυρίως μέθοδοι ανεξάρτητες καλλιέργειας, οι οποίες δεν βασίζονται στην ανάπτυξη μικροοργανισμών σε καλλιέργειες. Με αυτόν τον τρόπο παραδίδουν ολοκληρωμένη και αμερόληπτη αναφορά των οργανισμών παρουσών σε ένα δείγμα. Είναι οι προσδιοριστικές μέθοδοι που προσφέρουν την περισσότερη ευκαμψία και παραδίδουν τα πιο κατανοητά αποτελέσματα.

Αυτές οι μέθοδοι έγιναν δυνατές με την ανάπτυξη των τεχνικών PCR, οι οποίες πολλαπλασιάζουν συγκεκριμένα μέρη των DNA, με σκοπό να εμβαθύνουν την κάλυψη αυτών των περιοχών για αλληλούχιση, λειτουργική ανάλυση ενός συγκεκριμένου γονιδίου ή ανίχνευση πολυμορφισμών ή σημειακών μεταλλάξεων. Παράλληλα με άλλες εφαρμογές, η PCR είναι ικανή να ενισχύσει το 16S rRNA μικροοργανισμών, το οποίο είναι ο κυρίαρχος δείκτης ερευνών μικροοργανισμών, που απομονώνει το μεταλλακτικό-πληροφοριακό μέρος του DNA για αλληλούχιση. Έτσι η PCR επιτρέπει την μελέτη και αναγνώριση μικροοργανισμών που οι ερευνητές δεν κατάφεραν να καλλιεργήσουν ακόμα και αν το αρχικό δείγμα DNA ήταν μικρό.

1.4 Προσεγγίσεις NGS στην μεταγονιδιωματική

Όπως κάθε άλλη βιοπληροφορική ανάλυση, η πληροφορία που χρησιμοποιείται στη μεταγονιδιωματική/κλινική μεταγονιδιωματική προέρχεται από διαχείριση της πληροφορίας σε εργαστηριακό επίπεδο. Ακολουθείται ως γενικός κανόνας μια σειρά βημάτων για να δοθούν ως αποτέλεσμα δεδομένα που μπορούν να χρησιμοποιηθούν σε βιοπληροφορικές μεταγονιδιωματικές αναλύσεις. Η διαδικασία της μεταγονιδιωματικής ανάλυσης από τη λήψη δείγματος ως την βιοπληροφορική ανάλυση είναι επιγραμματικά : η συλλογή και η αποθήκευση του κλινικού δείγματος, η απομόνωση του DNA, η κατασκευή της γονιδιωματικής βιβλιοθήκης (κατάλληλη για

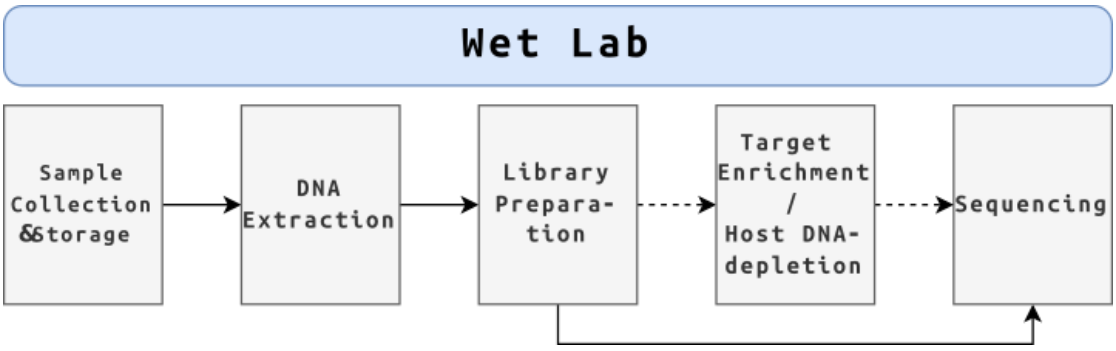
μεταγονιδιωματικές αναλύσεις) και η αλληλούχιση η οποία θα δίνει ως αποτέλεσμα ένα αρχείο fastq.

Αρχικά βέβαια είναι η λήψη του δείγματος που θα τεθεί προς ανάλυση από τον ασθενή. Τα δείγματα έχουν τη δυνατότητα να λάβουν οποιαδήποτε μορφή υγρού (αίμα, ούρα), στερεού (κόπρανα) ή άλλης μέθοδου (μάκτρο), όλα τα οποία έχουν λαμβάνονται από το ανθρώπινο σώμα (Loeffler et al., 2019). Αφού συλλεχθούν τα δείγματα, είναι απαραίτητη η σωστή αποθήκευσή τους σε συνθήκες κατάλληλες για τη διατήρησή τους. Έπειτα απαιτείται η απομόνωση του νουκλεϊκού οξέος, το οποίο θα χρησιμοποιηθεί στη συνέχεια στην αλληλούχιση. Το DNA και το RNA μπορούν να απομονωθούν είτε ταυτόχρονα είτε ξεχωριστά και η κατάλληλη μέθοδος θα αποφασιστεί ανάλογα με τις ανάγκες του στόχου της ανάλυσης. Μετά την απομόνωση, το επόμενο απαραίτητο βήμα είναι η προετοιμασία της βιβλιοθήκης (library preparation) για την αλληλούχιση, όπου τα ακατέργαστα νουκλεϊκά οξέα μετατρέπονται σε υλικό έτοιμο για αλληλούχιση (d'Humières et al., 2021). Υπάρχουν δύο βασικές προσεγγίσεις για την δημιουργία μιας NGS βιβλιοθήκης, η προετοιμασία με βάση την ένωση και η προετοιμασία με βάση την ετικετοποίηση. Σε κοινό επίπεδο και οι δύο μέθοδοι περιέχουν θρυμματισμό των δειγμάτων σε ισομεγέθη κομμάτια, διότι δεν είναι δυνατόν οι αλληλουχητές να αναλύσουν αλληλουχίες πλήρους μήκους. Μετά τον κατακερματισμό αυτό, τα θραύσματα DNA επισκευάζονται ή 'γυαλίζονται' στα άκρα τους. Συνήθως, μια μονή βάση αδενίνης προστίθεται για να σχηματιστεί μια προεξοχή μέσω μιας αντίδρασης Α-ουράς. Αυτή η προεξοχή Α επιτρέπει σε προσαρμογείς που περιέχουν μια ενιαία προεξέχουσα βάση θυμίνης να ζευγαρώσουν με τα θραύσματα DNA. Στη συνέχεια, ένα ένζυμο λιγάσης συνδέει ομοιοπολικά τον προσαρμογέα και εισάγει θραύσματα DNA, δημιουργώντας ένα πλήρες μόριο βιβλιοθήκης (López-Labrador et al., 2021). Αυτοί οι προσαρμογείς εξυπηρετούν πολλαπλές λειτουργίες. Μπορούν να περιλαμβάνουν γραμμωτούς κώδικες, που ονομάζονται επίσης ευρετήρια, για να αναγνωρίζουν δείγματα και να επιτρέπουν την πολυπλεξία. Υπάρχει τελικά η επιλογή ενίσχυσης της βιβλιοθήκης με PCR. Ο καθαρισμός του PCR προϊόντος μπορεί να πραγματοποιηθεί χρησιμοποιώντας μαγνητικά σφαιρίδια ή στήλη περιστροφής. Η βασική διαφορά της ένωσης με την ετικετοποίηση είναι η σειρά ολοκλήρωσης αυτών των βημάτων. Στη ένωση ο κατακερματισμός και η ένωση είναι δύο διαφορετικά βήματα που ολοκληρώνονται το ένα μετά το άλλο, ενώ στη ετικετοποίηση συμβαίνουν ταυτόχρονα.

Ένα προαιρετικό βήμα είναι η μείωση του DNA του ατόμου του δείγματος, διότι σε μεγάλες ποσότητες μπορεί να επισκιάσει αλληλουχίες που χρειάζονται στην βιοπληροφορική ανάλυση στη συνέχεια. Αυτό μπορεί να έχει ως αποτέλεσμα την απώλεια ορισμένου μικροβιακού ή ιικού υλικού. Κοινές μέθοδοι για την ολοκλήρωση της μείωσης είναι η αφαίρεση ολόκληρων κυττάρων και ο καθαρισμός των σωματιδίων (López-Labrador et al., 2021). Εναλλακτικά υπάρχει η διαδικασία μικροβιακού εμπλουτισμού, που αυξάνει το μικροβιακό DNA στο δείγμα, αυτό μπορεί να γίνει μετά την απομόνωση του νουκλεϊκού οξέος, ενώ η μείωση του DNA γίνεται πριν. Με τον μικροβιακό εμπλουτισμό μόνο ένα μέρος του γονιδιώματος εμπλουτίζεται και αλληλουχείται, ή περιοχές ενδιαφέροντος, χωρίς να αλληλουχηθεί όλο το γονιδίωμα ενός δείγματος. Υπάρχουν δύο βασικές μέθοδοι που χρησιμοποιούνται για αυτό το σκοπό στα πλαίσια της NGS, η αλληλούχιση amplicon και ο εμπλουτισμός στόχου υβριδικής σύλληψης. Η αλληλούχιση amplicon είναι μια ταχύτερη, ευκολότερη και πιο οικονομική επιλογή από τις εναλλακτικές που βασίζονται στον υβριδισμό. Ως εκ τούτου, ο προσδιορισμός της αλληλουχίας του amplicon είναι μια πιο κατάλληλη

επιλογή για αποτελεσματικές δοκιμές γονιδιακού πάνελ και εφαρμογές σε κλίμακα παραγωγής (συμπεριλαμβανομένης της κλινικής διάγνωσης και του βιομηχανικού ελέγχου γονιδιώματος). Επιπλέον, η αλληλουχία με αμπλικόνιο επιτρέπει την αντιμετώπιση χαμηλής εισαγωγής DNA, όπως η ευαισθησία ανίχνευσης και ανίχνευσης ιών και οι απαιτήσεις εισαγωγής DNA/RNA. Στον εμπλουτισμό στόχου υβριδικής σύλληψης, διερευνώνται σημαντικά μεγάλες περιοχές-στόχους, καθιστώντας το μια καλή επιλογή για έργα έρευνας και ανακάλυψης ευρύτερου εύρους. Θα πρέπει να σημειωθεί ότι αυτή η μέθοδος τείνει να έχει χαμηλό ρυθμό στόχου σε μικρότερα πάνελ λόγω της εγγενούς χαμηλότερης εξειδίκευσης των ανιχνευτών υβριδισμού. Για τους παραπάνω λόγους, οι επιστήμονες και οι κλινικοί ιατροί τείνουν να χρησιμοποιούν πολλαπλή PCR για μικρότερα πάνελ ή πάνελ hotspot για την ανίχνευση πολυμορφισμών μεμονωμένων νουκλεοτιδίων (SNPs) και/ή μικρών εισαγωγών/διαγραφών. Η υβριδική σύλληψη χρησιμοποιείται συνήθως για μεγάλες περιοχές-στόχους ή στόχους που περιλαμβάνουν σημαντικές δομικές αλλαγές, οι οποίες διαφορετικά θα ήταν πιο δύσκολο να στοχευθούν με μεθόδους που βασίζονται στην PCR.

Τέλος αρχίζει η αλληλούχιση του νουκλεϊκού οξέος που έχει απομονωθεί. Όπως έχει αναφερθεί προηγουμένως, στη μεταγονιδιωματική κυριαρχεί η αλληλούχιση NGS με βασικές πλατφόρμες την Illumina, ThermoFisher, BGI, Nanopore και PacBio. Η Illumina, BGI και PacBio καθορίζουν τα ζεύγη βάσεων χρησιμοποιώντας σημασμένα νουκλεοτίδια και taq πολυμεράσες. Το Nanopore τα καθορίζει περνώντας ένα μονόκλωνο DNA μέσα από ένα νανόπορο, η μοναδική διακοπή στη ροή ιόντων μέσα από τον νανοπόρο διαφοροποιείται ανά ομάδα νουκλεοτιδίων στο πόρο και το ThermoFisher χρησιμοποιεί επίσης τις αλλαγές του pH καθώς οι βάσεις ενσωματώνονται στην αλληλουχία (Loeffler et al., 2019). Γενική περιγραφή των μεθόδων αλληλούχισης έχει δοθεί σε παραπάνω κομμάτι θεωρίας αναφερόμενο στην NGS.



Εικόνα 1. Ροή εργαστηριακής εργασίας από την παραλαβή δείγματος μέχρι την αλληλούχιση. Με τη σειρά είναι : η συλλογή και αποθήκευση δείγματος, η εξαγωγή DNA, η προετοιμασία βιβλιοθήκης, προεραιτικά η αύξηση του οργανισμού στόχου ή η μείωση του DNA του ατόμου και τέλος η αλληλούχιση

Ως αποτέλεσμα της αλληλούχισης, δίνονται αρχεία fastq τα οποία χρησιμοποιούνται για τις διεργασίες στη συνέχεια. Μετά την ολοκλήρωση της αλληλούχισης, αρχίζει η βιοπληροφορική ανάλυση με την οποία ασχολείται κυρίως αυτή η παρούσα εργασία.

1.5 Σύγκριση Μεταγονιδιωματικών Εργαλείων

Παρόλο που η μεταγονιδιωματική είναι σχετικά νέα επιστήμη, υπάρχει ήδη πληθώρα εργαλείων. Η σύγκριση αυτών των εργαλείων μπορεί να υπάρξει δύσκολη λόγω των ποικίλων στρατηγικών εκτίμησης, συνόλων δεδομένων αναφοράς και κριτηρίων επίδοσης (Sczyrba et al., 2017). Επιπλέον λόγω της γρήγορης ανάπτυξης των εργαλείων και των τεχνολογιών η εκτίμηση τους απαιτεί σημαντικά ποσά χρόνου και υπολογιστικών πόρων.

Για την κατάλληλη σύγκριση αυτών των εργαλείων, που αλλιώς ονομάζονται ταξινομητές, είναι απαραίτητο να κατανοηθεί πόσο διαφορετικά λειτουργούν μεταξύ τους και πως να εκτιμηθεί η κατάλληλη προσέγγιση ανάλογα του είδους του δείγματος, του μικροβιακού βασιλείου ή της εφαρμογής. Για αυτό τον σκοπό απαιτείται η εξέταση τριών στοιχείων: ακρίβεια ταξινόμησης, ταχύτητα και υπολογιστικές απαιτήσεις (Ye et al., 2019). Στη συνέχεια, παρατίθεται σύγκριση των 5 πιο διαδεδομένων εργαλείων προσαρμοσμένη από άρθρο του Ye και των συνεργατών του, χρησιμοποιώντας κοινή βάση δεδομένων για να υπάρξει όσο μικρότερη διαφορά γίνεται μεταξύ τους.

Οι ταξινομητές που έχουν αναπτυχθεί απαιτούν προ-διαμορφωμένες βάσεις δεδομένων που περιέχουν αλληλουχίες στις οποίες θα αντιστοιχηθούν τα δεδομένα που εισάγονται. Υπάρχουν δύο βασικά είδη εργαλείων που όμως μπορούν πρακτικά να χρησιμοποιηθούν για τον ίδιο σκοπό εναλλακτικά μεταξύ τους χωρίς κάποιο πρόβλημα. Αυτά είναι ως εξής το ταξινομικό ‘binning’ και το ταξινομικό ‘profiling’. Το ‘binning’ περιλαμβάνει την κατάταξη μοναδικών αλληλουχιών σε ταξινομία αναφοράς. Το ‘profiling’ αναφέρει τον σχετικό πλήθος ταξινομιών σε ένα δείγμα αλλά δε κατατάσει μοναδικές αλληλουχίες. Ένα ταξινομικό προφίλ μπορεί να υπολογισθεί από προσέγγιση ‘binning’ ως το άθροισμα των ταξινομικών κατατάξεων που δίνονται ως αποτέλεσμα.

Οι ταξινομητές χρησιμοποιούν μεθόδους για να εξασφαλίσουν την γρήγορη παραγωγή αποτελεσμάτων. Αυτές περιλαμβάνουν την δημιουργία μικρών ομάδων βάσεων όπως τα k-mers αντί τον έλεγχο κάθε ξεχωριστής βάσης ή την χρήση FM index (ένα συμπιεσμένο ευρετήριο υποσυμβολοσειράς πλήρους κειμένου με βάση τον μετασχηματισμό Burrows–Wheeler, με κάποιες ομοιότητες με τον πίνακα επιθημάτων) και την χρήση περισσότερης μνήμης με σκοπό την μείωση χρήσης μεγάλου ποσού CPU. Τα εργαλεία αυτά μπορούν να διαχωριστούν σε τρεις κατηγορίες ταξινομήσεων : DNA-to-DNA, DNA-to-protein και βασισμένες σε δείκτες. Η συγκεκριμένη σύγκριση θα επικεντρωθεί σε εργαλεία ταξινόμησης DNA-to-DNA τα οποία συγκρίνουν τις αλληλουχίες με γενομικές βάσεις δεδομένων DNA και RNA (Ye et al., 2019).

Οι πιο κοινά χρησιμοποιημένες βάσεις δεδομένων για τη μεταγονιδιωματική είναι η SILVA (Quast et al., 2013), η Greengenes (DeSantis et al., 2006), η RDP (Cole et al., 2014), η RefSeq (O’Leary et al., 2016) και η BLAST nt (Morgulis et al., 2008). Οι τρεις πρώτες εξειδικεύονται σε αλληλουχίες γονιδιωμάτων 16S RNA, η RefSeq (O’Leary et al., 2016) περιέχει ολοκληρωμένα γονιδιώματα για είδη μικροβίων και η BLAST nt (Morgulis et al., 2008) έχει νουκλεοτιδικές αλληλουχίες υψηλής ποιότητας. Συνήθως το μέγεθος αυτών των βάσεων δεδομένων φτάνει τα 50-100GB και συνεχώς ανανεώνονται λόγω του γρήγορου ρυθμού προόδου στην κατηγοριοποίηση πρόσφατα ανιχνευμένων οργανισμών.

Όσον αφορά τις μετρήσεις που θα εξεταστούν , οι πιο σημαντικές στην ταξινόμηση είναι η ακρίβεια και η ευαισθησία. Ως ακρίβεια θεωρείται το ποσοστό των αληθινά θετικών ειδών ανιχνευμένων στο δείγμα δια του συνολικού αριθμού ειδών που ανιχνεύθηκαν μέσω της μεθόδου. Ενώ η ευαισθησία είναι το ποσοστό των αληθινά θετικά αναγνωρισμένων ειδών διά του συνολικού αριθμού ξεχωριστών ειδών που υπάρχουν πραγματικά στο δείγμα. Επίσης εξετάζεται πόσο χώρο λαμβάνει στη μνήμη, η ακρίβεια με την οποία έχουν ανιχνευθεί οι ποσότητες κάθε είδους ή γένους σε σχέση με το αρχικό δείγμα, η μορφή αρχείου της εξόδου, η ταχύτητα και ο χρόνος ολοκλήρωσης της ανάλυσης.

Για την εκτίμηση της ακρίβειας, υπολογίζονται οι ζευγαρωμένες αποστάσεις μεταξύ αλήθειας και οι κανονικοποιημένες αφθονίες για κάθε προσδιορισμένη ταξινόμηση σε ένα δεδομένο ταξινομικό επίπεδο.

Τα εργαλεία ταξινόμησης που εκτιμήθηκαν ήταν τα Kraken (Wood and Salzberg, 2014), Kraken 2 (Wood et al., 2019), CLARK (Ounit et al., 2015), Centrifuge (Kim et al., 2016) και PathSeq (Walker et al., 2018). Η βιβλιοθήκη που χρησιμοποιήθηκε ήταν βασισμένη στο RefSeq CG (O'Leary et al., 2016) και εξετάστηκε έναντι των αντίστοιχων προκαθορισμένων βάσεων για κάθε εργαλείο. Για να ελεγχθεί αρχικά η ακρίβεια και η ευαισθησία των εργαλείων 'έτρεξαν' χρησιμοποιώντας 10 σύνολα δεδομένων αναφοράς τα οποία ήταν υπολογιστικά προσομοιωμένα από 12 έως 525 βακτηριακά είδη και αναπτύχθηκε αναφορά των αναγνωρισμένων ταξινομιών και οι αντίστοιχες ποσότητές τους μέσα στο δείγμα.

Σημειώθηκε μια σημαντική παρατήρηση κατά τις αναλύσεις. Αρχικά, η προεπιλεγμένη βάση δεδομένων του Centrifuge (Kim et al., 2016) συμπιέζεται σημαντικά με απώλειες (εναποθέτοντας τις ακολουθίες βάσης δεδομένων για εξοικονόμηση χώρου), γεγονός που οδηγεί επίσης σε κακή ευαισθησία. Ως αποτέλεσμα, σε σύγκριση των αρχικών αποτελεσμάτων με τα αποτελέσματα που χρησιμοποιούν την κοινή βάση δεδομένων, τα δεύτερα έδειξαν βελτιωμένο βαθμό λόγω της καλύτερης ποιότητας βάσης.

Τελικά βρέθηκε πως εργαλεία που χρησιμοποιούσαν k-mers, συμβολοσειρές μήκους k που περιέχονται σε μια βιολογική αλληλουχία, με μήκος άνω των 30 ζβ, όπως το Kraken 2 (Wood et al., 2019), είχαν τα καλύτερα αποτελέσματα, με απόκλιση από την αλήθεια κάτω του 0.1%. Οι υπολογιστικές ανάγκες των προγραμμάτων μπορεί να υπάρξουν απαγορευτικές για ορισμένες εφαρμογές όμως η χρήση λιγότερο απαιτητικών προγραμμάτων θα έχει ως συνέπεια λιγότερο ακριβή αποτελέσματα. Ενώ τα περισσότερα προγράμματα είχαν εξόδους παρόμοιας ποιότητας, το Kraken 2 (Wood et al., 2019) ξεχώρισε λαμβάνοντας λιγότερο χώρο από τα υπόλοιπα και δίνοντας τα αποτελέσματα σε μικρότερο χρόνο.

Classifier	Custom Databases	Generates Abundance Profile	Memory Required (in Gb)	Time Required (in minutes)	Reference
Centrifuge	yes	yes	20	7	Kim et al., 2016
CLARK	yes	yes	80	2	Ounit et al., 2015
Kraken	yes	yes	190	1	Wood and Salzberg, 2014
Kraken 2	yes	yes	36	1	Wood et al., 2019
PathSeq	no	yes	140	5	Walker et al., 2018

Πίνακας 1. Πίνακας με τα εργαλεία και πληροφορίες γι' αυτά, συγκεκριμένα : όνομα, αν υποστηρίζουν βάση δεδομένων δημιουργημένη από τον χρήστη, πόση μνήμη απαιτείται, χρόνος που απαιτείται για την ολοκλήρωση της ανάλυσης και παραπομπή στο άρθρο του εργαλείου Προσαρμοσμένος από (Ye et al., 2019)

1.6 Σκοπός

Σκοπός της παρούσας πτυχιακής εργασίας είναι η ανάπτυξη βιοπληροφορικών ροών μεταγονιδιωματικής ανάλυσης κλινικών δεδομένων που θα ενσωματωθούν σε διαδραστική πλατφόρμα που θα κατασκευαστεί με στόχο την δημοσία προσβασιμότητα και φιλοξενία εργαλείων και βιοπληροφορικών ροών για την ανάλυση high throughput κλινικών δεδομένων επόμενης γενιάς (NGS). Ο λόγος προσέγγισης κλινικών δεδομένων είναι η υπάρχουσα ανάγκη για βελτίωση των τεχνολογιών γύρω από το αντικείμενο αυτό. Οι βιοπληροφορικές ροές είναι βασισμένες σε πληθώρα βασικών λειτουργιών που είναι απαραίτητες σε μία πλήρη μεταγονιδιωματική ανάλυση, όπως η ταξινόμηση και η ανάλυση ποικιλότητας. Στην παρούσα κατάσταση, η πλατφόρμα έχει δυνατότητα λειτουργίας σε τοπικό επίπεδο, αλλά ο στόχος χρήσης της δημόσια στον παγκόσμιο ιστό, φιλοξενώντας όχι μόνο ήδη υπάρχουσα εργαλεία και ροές αλλά και μελλοντικά ανεπτυγμένα. Σημαντική διαφορά με την αρχική λειτουργία που προσφέρει το Galaxy (Afgan et al., 2018) είναι η προκαθορισμένες βέλτιστες ρυθμίσεις των εργαλείων, η ύπαρξη συγκεκριμένων μόνο εργαλείων για αναλύσεις περιορίζοντας έτσι το χάος επιλογών για άπειρους χρήστες, η δυνατότητα ενημέρωσης των χρηστών ως προς την ολοκλήρωση της ανάλυσής τους μέσω email, η χρηστική διεπαφή η οποία διευκολύνει τη χρήση των εργαλείων και ροών ανάλυσης και στο μέλλον, η δοκιμή νέων εργαλείων τα οποία δεν διατίθενται στο Galaxy (Afgan et al., 2018).

2. Μεθοδολογία

2.1 Προγραμματιστικά Εργαλεία

Τα βασικά εργαλεία που χρησιμοποιήθηκαν στα πλαίσια της παρούσας πτυχιακής εργασίας είναι η προγραμματιστική γλώσσα Python σε συνδυασμό με τη δομή Django και την ανοιχτή πλατφόρμα Galaxy (Afgan et al., 2018) για την ανάλυση δεδομένων σε συνδυασμό με το BioBlend (Sloggett et al., 2013).

2.1.1 Python

Η γλώσσα προγραμματισμού Python αποτελεί την πλέον προτιμώμενη γλώσσα προγραμματισμού εφαρμογών διαδικτύου, καθώς συνδυάζει απλότητας γραφής και ευκολία εκμάθησης. Όντας μια ευέλικτη γλώσσα, προσφέρει γρήγορη ανάπτυξη εφαρμογών που βασίζονται στο διαδίκτυο. Προσφέρει ανάπτυξη χρησιμοποιώντας WSGI (περιγράφει την σύνδεση διακομιστή ιστού με εφαρμογές ιστού και πώς αυτές οι εφαρμογές μπορούν να συνδεθούν για να εκπληρώσουν ένα αίτημα) και ASGI (διάδοχος του WSGI το οποίο διαθέτει διεπαφή σε σύγχρονες και ασύγχρονες εφαρμογές, πλαίσια και διακομιστές ιστού) . Η γλώσσα προγραμματισμού Python επιλέχθηκε έναντι της Java, PHP ή Ruby καθώς δίνει την δυνατότητα εφαρμογής απλής διεπαφής μεταξύ του διακομιστή και του χρήστη και την απαλλαγή από την ανάγκη προγραμματισμού λειτουργικών θεμελίων της εφαρμογής τα οποία ήδη έχει καλύψει η γλώσσα και απαλλάσσει τον χρήστη από την σπατάλη χρόνου και υπολογιστικών πόρων που απαιτούν οι προαναφερόμενες γλώσσες προγραμματισμού, επιτρέποντας του να «αφοσιωθεί» στην εκτέλεση πιο κρίσιμων λειτουργιών και παραμέτρων. Έχει ενσωματωμένες δομές δεδομένων όπως λίστες και πίνακες κατακερματισμού (dictionaries) και μια αρκετά μεγάλη βιβλιοθήκη.

2.1.2 Django

Η δομή Django παρέχει βιβλιοθήκες και ενότητες οι οποίες απλοποιούν διαδικασίες οι οποίες σχετίζονται με την διαχείριση πληροφοριών, την αλληλεπίδραση με τη βάση δεδομένων και διασύνδεση με διαφορετικά πρωτόκολλα διαδικτύου όπως HTTP, SMTP, XML-RPC και FTP. Ως εφαρμογή διαδικτύου, η ασφάλεια των πληροφοριών είναι της ύψιστη σημασίας γι’αυτό το λόγο η Django χρησιμοποιεί τεχνικές όπως clickjacking, cross-site scripting και SQL injections για να ‘οχυρώσει’ την εφαρμογή.

Το διαδίκτυο λειτουργεί μέσω της επικοινωνίας εξυπηρετητών και χρηστών. Η επικοινωνία αυτή πραγματοποιείται με την αποστολή και απάντηση αιτημάτων. Η κύρια ευθύνη της Django ως δομή είναι ο έλεγχος των αιτημάτων των χρηστών της πλατφόρμας. Κάθε φορά που ένα αίτημα χρήστη αποσταλεί στον διακομιστή δικτύου αυτό μεταφέρεται στην Django η οποία ελέγχει την διεύθυνση που ζητείται και αν το αίτημα είναι εφικτό ή όχι ανάλογα με τις πληροφορίες που έχουν δοθεί σε σχέση με τη λειτουργία view της εφαρμογής και απαντάει ανάλογα, είτε με την πραγματοποίηση του αιτήματος, είτε με την απόρριψη του και την εμφάνιση μηνύματος σφάλματος.

2.1.3 Galaxy

Το πλαίσιο Galaxy (Afgan et al., 2018) ενσωματώνει υπολογιστικά εργαλεία προηγμένης τεχνολογίας και παρέχει εύχρηστες διεπαφές χρήστη, ενώ κρύβει τις λεπτομέρειες της διαχείρισης υπολογιστών και αποθήκευσης. Εξαλείφει έτσι την ανάγκη για εξειδικευμένη τεχνογνωσία στην πληροφορική κατά την εκτέλεση πολλών κοινών τύπων ανάλυσης μεγάλης κλίμακας. Επιτρέπει στους χρήστες να εκτελούν υπολογιστικές αναλύσεις ποικίλων αρχείων δεδομένων, συμπεριλαμβανομένων γονιδιωματικών και μεταγονιδιωματικών δεδομένων. Η δημόσια υπηρεσία Galaxy (Afgan et al., 2018) καθιστά διαθέσιμα εργαλεία ανάλυσης, γονιδιωματικά δεδομένα, επιδείξεις εκμάθησης, μόνιμους χώρους εργασίας και υπηρεσίες δημοσίευσης σε κάθε επιστήμονα που έχει πρόσβαση στο Διαδίκτυο. Το Galaxy (Afgan et al., 2018) έχει δημιουργήσει μια σημαντική κοινότητα χρηστών και προγραμματιστών.

Επιτρέπει στους χρήστες να εκτελούν ολοκληρωμένες γονιδιωματικές αναλύσεις παρέχοντας μια ενοποιημένη, βασισμένη στον ιστό διεπαφή για τη λήψη γονιδιωματικών δεδομένων και την εφαρμογή υπολογιστικών εργαλείων για την ανάλυση των δεδομένων. Οι χρήστες μπορούν να εισάγουν σύνολα δεδομένων στους χώρους εργασίας τους από πολλές καθιερωμένες αποθήκες δεδομένων (πχ Google drive) ή να ανεβάζουν τα δικά τους σύνολα δεδομένων στον αποθηκευτικό χώρο που παρέχεται στον κάθε χρήστη. Οι διεπαφές στα υπολογιστικά εργαλεία δημιουργούνται αυτόματα από αφηρημένες περιγραφές για να διασφαλιστεί μια συνεπής εμφάνιση και αίσθηση.

Παράλληλα, τα επιστημονικά συστήματα ροής εργασιών έχουν φτάσει σε ένα επίπεδο ωριμότητας που τα καθιστά βολικά για τον προγραμματισμό της εκτέλεσης πολύπλοκων και μεγάλης κλίμακας αναλύσεων, ενώ διαχειρίζονται σωστά τα δεδομένα παρακολουθώντας δεδομένα που καταναλώνονται και παράγονται. Για τους πειραματιστές, παρέχει ένα περιβάλλον ανάλυσης στο οποίο μπορούν να εκτελούν ανάλυση διαδραστικά, διασφαλίζοντας παράλληλα ότι οι αναλύσεις που προκύπτουν είναι διαφανείς και αναπαραγώγιμες.

Υπάρχουν ορισμένες ορολογίες και έννοιες οι οποίες απαιτούνται για την κατανόηση των διεργασιών της πλατφόρμας. Πολλές από τους τύπους δεδομένων που χρησιμοποιούνται στη γονιδιωματική αποτελούνται από σειρές στηλών οριοθετημένης καρτέλας που περιέχουν ποικίλα δεδομένα. Ένα από αυτά, γνωστό ως αρχείο BED, στο οποίο κάθε σειρά αντιπροσωπεύει τη θέση ενός γονιδιωματικού χαρακτηριστικού σε ένα συγκεκριμένο γονιδίωμα.Το BED αρχείο περιέχει τουλάχιστον τρεις στήλες: (1) το χρωμόσωμα, (2) τη θέση έναρξης εντός αυτού του χρωμοσώματος και (3) την τελική θέση εντός αυτού του χρωμοσώματος. Άλλες στήλες που περιλαμβάνονται συνήθως είναι πληροφορίες ονόματος, κλώνου, βαθμολογίας και εξονίου (όταν τα διαστήματα είναι σχολιασμοί γονιδίων). Χρησιμοποιούνται πρόσθετες μορφές πέρα από αυτές που αποτελούνται από στήλες πίνακα, αλλά οι επιπλοκές των μορφών τους μπορούν να αγνοηθούν σε μεγάλο βαθμό σε αυτό το εισαγωγικό κείμενο, καθώς το Galaxy (Afgan et al., 2018) μπορεί να χειριστεί τις περισσότερες από τις λεπτομέρειες που απαιτούνται για την εκτέλεση σύνθετης ανάλυσης. Κάθε φορά που εκτελείται ένα εργαλείο, δημιουργούνται ένα ή περισσότερα σύνολα δεδομένων στο ιστορικό του χρήστη. Όταν εκτελείται ανάλυση στο Galaxy (Afgan et al., 2018), κάθε λεπτομέρεια διατηρείται στο «ιστορικό» και μπορεί να αναθεωρηθεί αργότερα. Αυτά τα ιστορικά μπορούν να κοινοποιηθούν ή να δημοσιευτούν και μπορούν να αναπαραχθούν (με ή χωρίς τροποποίηση) μέσω του συστήματος ροής εργασιών. Έτσι, χωρίς πρόσθετη προσπάθεια εκ μέρους του χρήστη, το Galaxy (Afgan et al., 2018) διευκολύνει τη

μεγαλύτερη διαφάνεια και αναπαραγωγιμότητα των υπολογιστικών αναλύσεων. Όταν ένα σύνολο δεδομένων είναι έτοιμο για προβολή ή χρήση παρατίθενται πρόσθετα βήματα ανάλυσης. Τα εργαλεία που επιλέγονται προς χρήση περιμένουν στην ουρά για να ολοκληρωθεί το σύνολο δεδομένων που απαιτείται πριν από την εκτέλεση.

2.1.4 BioBlend

Το Galaxy (Afgan et al., 2018) δίνει μια γραφική διεπαφή χρήστη (GUI) για την εξακρίβωση της πληροφορίας η οποία θα επεξεργαστεί, ποια βήματα θα ακολουθηθούν και την σειρά εκτέλεσής τους. Ωστόσο, όσο βολικά και φιλικά προς το χρήστη κι ας είναι, τα GUI δεν είναι κατάλληλα για αυτοματοποιημένη ανάλυση και μαζική επεξεργασία. Για παράδειγμα, όταν γίνεται διαθέσιμη μια νέα έκδοση ενός γονιδιώματος αναφοράς ή όταν ενημερωθεί η έκδοση ενός εργαλείου, για να διασφαλιστεί ότι τα αποτελέσματα της ανάλυσης παραμένουν σχετικά, απαιτείται το πλήρες σύνολο πειραματικών αποτελεσμάτων (π.χ. ανακάλυψη πολυμορφισμού νουκλεοτιδίων) να επαναξιολογηθεί από την αρχή χρησιμοποιώντας τα δεδομένα του νέου μοντέλου ή το λογισμικό. Αυτή η επίπονη εργασία απαιτεί καλύτερη υποστήριξη από το υπολογιστικό πλαίσιο που χρησιμοποιείται, με τη μορφή αξιόπιστων τρόπων για την αυτοματοποίηση των λειτουργιών, τη μαζική επεξεργασία συνόλων δεδομένων και την τεκμηρίωση της ανάλυσης που εκτελείται σε οποιαδήποτε από αυτές. Γενικότερα, οι μελέτες τείνουν να χειρίζονται έναν αυξανόμενο αριθμό δειγμάτων. Τείνουν επίσης να διαρκούν περισσότερο από τους σχετικά συχνούς κύκλους ενημέρωσης για δεδομένα μοντέλου και λογισμικό. Και οι δύο αυτές συνθήκες θέτουν απαιτήσεις για τέτοιες αυτοματοποιημένες λειτουργίες μαζικών δεδομένων που επί του παρόντος δεν αντιμετωπίζονται σωστά από τα GUI.

Για να διευκολυνθεί αυτού του είδους η επεξεργασία, το Galaxy (Afgan et al., 2018) περιλαμβάνει μια διεπαφή προγραμματισμού εφαρμογών RESTful (Richardson and Ruby, 2007) που επιτρέπει σε άλλα προγράμματα να την ελέγχουν αυτόματα. Ωστόσο, αυτό το API είναι αρκετά χαμηλού επιπέδου, καθώς απαιτεί από τους χρήστες να κατασκευάζουν και να εκδίδουν αιτήματα HTTP, να χειρίζονται ρητά τις τυπικές περιπτώσεις σφαλμάτων που εμφανίζονται σε τέτοια κατανεμημένα σενάρια και να φροντίζουν για τη σειριοποίηση και την αποσειριοποίηση δεδομένων στις ανταλλαγές μεταξύ του χρήστη και του διακομιστή. Αυτό το κενό στη λειτουργικότητα οδήγησε στην ανάπτυξη του BioBlend (Sloggett et al., 2013), ενός πακέτου Python που κρύβει την επικοινωνία HTTP, τον χειρισμό σφαλμάτων και την (απο)σειριοποίηση JSON από τον χρήστη, παρέχοντας ένα API βασισμένο σε λεξικό που απλοποιεί σημαντικά την αλληλεπίδραση με το Διακομιστής Galaxy (Afgan et al., 2018).

Επί του παρόντος, το API Galaxy (Afgan et al., 2018) είναι διαθέσιμο ως διεπαφή REST, η οποία είναι αναμφισβήτητα δύσκολη στην κατανόηση και χρήση δεδομένου ότι περισσότεροι βιοπληροφορικοί νιώθουν άνετα να γράφουν προγράμματα αυτοματισμού σε γλώσσα προγραμματισμού υψηλού επιπέδου. Έχοντας αυτό κατά νου, έχουν αναπτύξει μια βιβλιοθήκη Python για το Galaxy (Afgan et al., 2018), που ονομάζεται BioBlend (Sloggett et al., 2013) που παρέχει μια διεπαφή υψηλού επιπέδου για αλληλεπίδραση με τις δύο εφαρμογές. Αυτό προάγει την ταχύτερη αλληλεπίδραση με αυτές τις εφαρμογές, διευκολύνει την επαναχρησιμοποίηση και την κοινή χρήση σεναρίων και διευκολύνει τη συνεργασία μεταξύ βιοπληροφορικών και βιολόγων. Παρέχεται εκτεταμένη τεκμηρίωση API για τη βιβλιοθήκη, ενώ το αποθετήριο πηγαίου κώδικα περιέχει συγκεκριμένα παραδείγματα.

Παρά τις σημαντικές βελτιώσεις του σε σχέση με την ακατέργαστη διεπαφή χαμηλού επιπέδου, το BioBlend (Sloggett et al., 2013) έχει περιθώρια βελτίωσης. Για παράδειγμα, το μεγαλύτερο μέρος του BioBlend (Sloggett et al., 2013) API εξακολουθεί να προσφέρει μια αντιστοίχιση ενός προς ένα των γενικών λεξικών Python στους πόρους του Galaxy (Afgan et al., 2018) REST, χωρίς ρητή μοντελοποίηση των οντοτήτων Galaxy και των σχέσεών τους. Επίσης, η διεπαφή αποτυγχάνει να απομονώσει τον κώδικα του χρήστη από αλλαγές στο Galaxy (Afgan et al., 2018) API, καθώς μεταβιβάζει στον καλούντα τις ίδιες δομές λεξικού που στέλνει ο διακομιστής. Τέλος, το BioBlend (Sloggett et al., 2013) δεν έχει τη δυνατότητα παροχής εκτέλεσης λειτουργιών υψηλότερου επιπέδου παρά την προνομιακή θέση του παρέχει πολλά με την «πλούσια» λειτουργικότητα για την εκτέλεση εργασιών υψηλότερου επιπέδου, εξελιγμένες αλλά γενικές, παρά το γεγονός ότι είναι τοποθετημένο σε μια προνομιακή θέση στη στοίβα λογισμικού, όπου δυνητικά είναι κοινόχρηστο από όλες τις εφαρμογές-πελάτες του χρήστη.

Γι' αυτό αναπτύχθηκε το BioBlend.objects (Leo et al., 2014), μια διεπαφή Galaxy (Afgan et al., 2018) που υλοποιείται ως νέο επίπεδο πάνω από το BioBlend(Sloggett et al., 2013). Το νέο API αντιμετωπίζει τα προαναφερθέντα ζητήματα με δύο βασικά χαρακτηριστικά: ένα μοντέλο αντικειμενοστραφούς προγραμματισμού (Object Oriented - OO), το οποίο απλοποιεί την ανάπτυξη και απομονώνει τον κώδικα του χρήστη από αλλαγές στο Galaxy (Afgan et al., 2018) API και ένα στοιχείο υψηλού επιπέδου που απλοποιεί πολύπλοκες λειτουργίες και υποστηρίζει τον μετα υπολογισμό των πληροφοριών που περιγράφουν τις διάφορες οντότητες του Galaxy (Afgan et al., 2018). Με το BioBlend.objects, η εκτέλεση μιας ροής εργασίας Galaxy (Afgan et al., 2018) απαιτεί μόνο μερικές γραμμές απλού κώδικα.

Το BioBlend.objects έχει αναπτυχθεί ως υπομονάδα της αρχικής βιβλιοθήκης BioBlend (Sloggett et al., 2013). Ιεραρχικά, ο κώδικας βρίσκεται επί του παρόντος στο ίδιο επίπεδο με τις υπομονάδες Galaxy (Afgan et al., 2018) του BioBlend (Sloggett et al., 2013). Η βιβλιοθήκη αποτελείται από δύο κύρια στοιχεία: τη μονάδα wrappers, η οποία ορίζει τη δομή αντικειμένων που αντικατοπτρίζει τις οντότητες του Galaxy (Afgan et al., 2018) και τη μονάδα του χρήστη, ένα επίπεδο κώδικα υψηλού επιπέδου που βασίζεται στο αρχικό API για να εκθέσει μια απλούστερη, πιο συνοπτική διεπαφή με βάση το αντικείμενο ιεραρχία που ορίζεται σε περιτυλίγματα (wrappers). Η ενότητα χρήστη αποτελείται από τρεις κύριες κατηγορίες που ενσωματώνουν τις αλληλεπιδράσεις με τις πιο σημαντικές οντότητες του Galaxy (Afgan et al., 2018): ιστορικά, ροές εργασίας και βιβλιοθήκες. Η λειτουργική μονάδα galaxy_instance περιέχει την κλάση GalaxyInstance, η οποία ενοποιεί τους τρεις χρήστες, ενεργώντας ως κοινό σημείο εισόδου για όλες τις αλληλεπιδράσεις με τον διακομιστή Galaxy (Afgan et al., 2018).

Ως βιβλιοθήκη προγραμμάτων, το BioBlend (Sloggett et al., 2013) επιτρέπει την αυτοματοποίηση τόσο της επεξεργασίας αγωγών όσο και της παροχής και διαχείρισης υποδομής. Ως αποτέλεσμα, είναι μοναδικά τοποθετημένο για τον εξορθολογισμό της αυτοματοποίησης αγωγών και για να γίνει μια τυπική βιβλιοθήκη για αλληλεπίδραση με το Galaxy (Afgan et al., 2018). Στο τέλος μιας εκτέλεσης, η υπολογιστική υποδομή μπορεί να απελευθερωθεί αυτόματα, ενώ όλα τα δεδομένα και τα βήματα που εκτελούνται διατηρούνται στο Galaxy (Afgan et al., 2018), επιτρέποντας εύκολη επαναχρησιμοποίηση, οπτική αλληλεπίδραση, κοινή χρήση και περαιτέρω ανάλυση. Επιπλέον, είναι δυνατό να οραματιστούμε το BioBlend (Sloggett et al., 2013) ως ένα

πρώτο βήμα προς τον καθορισμό ενός Galaxy Shell, το οποίο θα επέτρεπε μια πιο ολοκληρωμένη πρόσβαση στα εσωτερικά του Galaxy (Afgan et al., 2018) για προηγμένη χρήση.

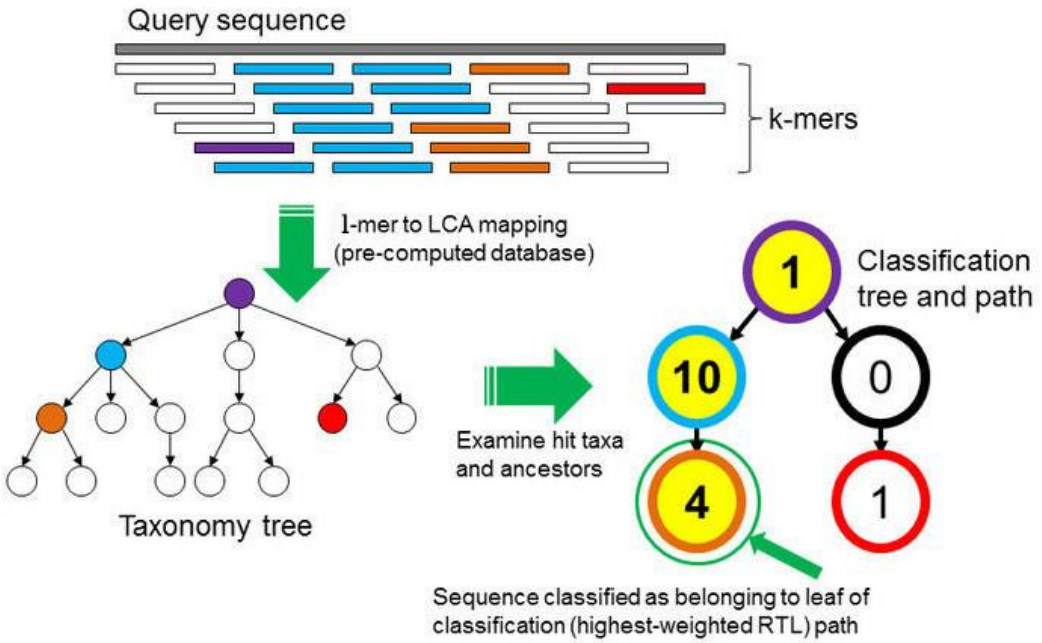
2.2 Μεταγονιδιωματικά Εργαλεία Ανάλυσης και Οπτικοποίησης

2.2.1 Kraken 2

Το πιο διαδεδομένα χρησιμοποιημένο εργαλείο για ανάλυση 16S rRNA είναι το QIIME (Kuczynski et al., 2011)/QIIME2 (Bolyen et al., 2019), το οποίο συγκρίνει αλληλουχίες έναντι βιβλιοθήκης αναφοράς 16S (συνήθως Greengenes (Quast et al., 2013), SILVA (Quast et al., 2013) ή RDP (Cole et al., 2014)). Το μειονέκτημα του είναι το πόσο υπολογιστικά δαπανηρό είναι, απαιτώντας αρκετά περισσότερο CPU και χώρο μνήμης από άλλα εργαλεία.

Το Kraken 2 (Wood et al., 2019) χρησιμοποιεί έναν ταχύτατο και ακριβή αλγόριθμο. Χρησιμοποιώντας ένα μόνο thread, το Kraken 2 (Wood et al., 2019) μπορεί να ταξινομήσει τα δεδομένα ακολουθίας με ρυθμό > 1 εκατομμύριο αναγνώσεις ανά λεπτό. Το Kraken 2 (Wood et al., 2019), \παρουσιάζει μεγάλες αλλαγές στη δόμηση βάσης δεδομένων και στα βήματα ταξινόμησης ώστε να γίνουν μικρότερες οι βάσεις και γρηγορότερες οι ταξινομήσεις. Το Kraken 2 (Wood et al., 2019) έχει σχεδόν την ίδια ακρίβεια και ευαισθησία με το Kraken (Wood and Salzberg, 2014) με διαφορά ότι προσφέρει υποστήριξη για την ταξινόμηση 16S rRNA με τις τρεις βασικές βάσεις δεδομένων.

Το Kraken 2 (Wood et al., 2019) είναι ένας ταξινομητής αλληλουχιών που εκχωρεί ταξινομικές σημάνσεις σε αλληλουχίες DNA. Το Kraken 2 (Wood et al., 2019, p. 2) εξετάζει τα k-mers (συμβολοσειρές μήκους k που περιέχονται σε μια βιολογική ακολουθία) μέσα σε μια ακολουθία και χρησιμοποιεί τις πληροφορίες μέσα σε αυτά τα k-mers για να απευθυνθεί σε μια βάση δεδομένων. Αυτή η βάση δεδομένων αντιστοιχεί τα k-mers στον χαμηλότερο κοινό πρόγονο (LCA) όλων των γονιδιωμάτων που είναι γνωστό ότι περιέχουν ένα δεδομένο k-mer. Δίνοντας μας ως αποτέλεσμα την ταξινόμηση κάθε αλληλουχίας στην πιο πιθανή ταξινομική βαθμίδα. Για να επιτευχθεί αυτό το αποτέλεσμα χρησιμοποιούνται διάφορες βάσεις δεδομένων, μία από αυτές είναι η MiniKraken η οποία χρειάζεται λιγότερη μνήμη και υπολογιστικό χώρο για τη χρήση της, αυτό μαζί με το γεγονός ότι είναι η καλύτερα διαμορφωμένη βάση για τις ανάγκες του εργαλείου είναι ο λόγος γιατί χρησιμοποιείται στη ροή.



Εικόνα 2. Διαγραμματική ροή εργαλείων Kraken 2 (Gallardo et al 2020, Batut et al., 2018)

Σε σύγκριση QIIME2 (Bolyen et al., 2019) και Kraken 2 (Wood et al., 2019) χρησιμοποιώντας τις βάσεις δεδομένων Greengenes (Quast et al., 2013) και SILVA (Quast et al., 2013) 16S rRNA. Το Kraken 2 (Wood et al., 2019) επιτρέπει την χρήση πολλαπλών νημάτων για να επιταχύνει την κατασκευή βάσης δεδομένων, ενώ το QIIME (Kuczynski et al., 2011) χρησιμοποιεί μόνο ένα νήμα. Το Kraken 2 (Wood et al., 2019) 'έχτισε' βάση δεδομένων σχεδόν εννιά φορές πιο γρήγορα από το QIIME2 (Bolyen et al., 2019). Η απαιτούμενη υπολογιστική μνήμη του Kraken 2 (Wood et al., 2019) ήταν πολύ μικρότερη του QIIME2 (Bolyen et al., 2019) (3,4 GB για Greengenes (Quast et al., 2013)).

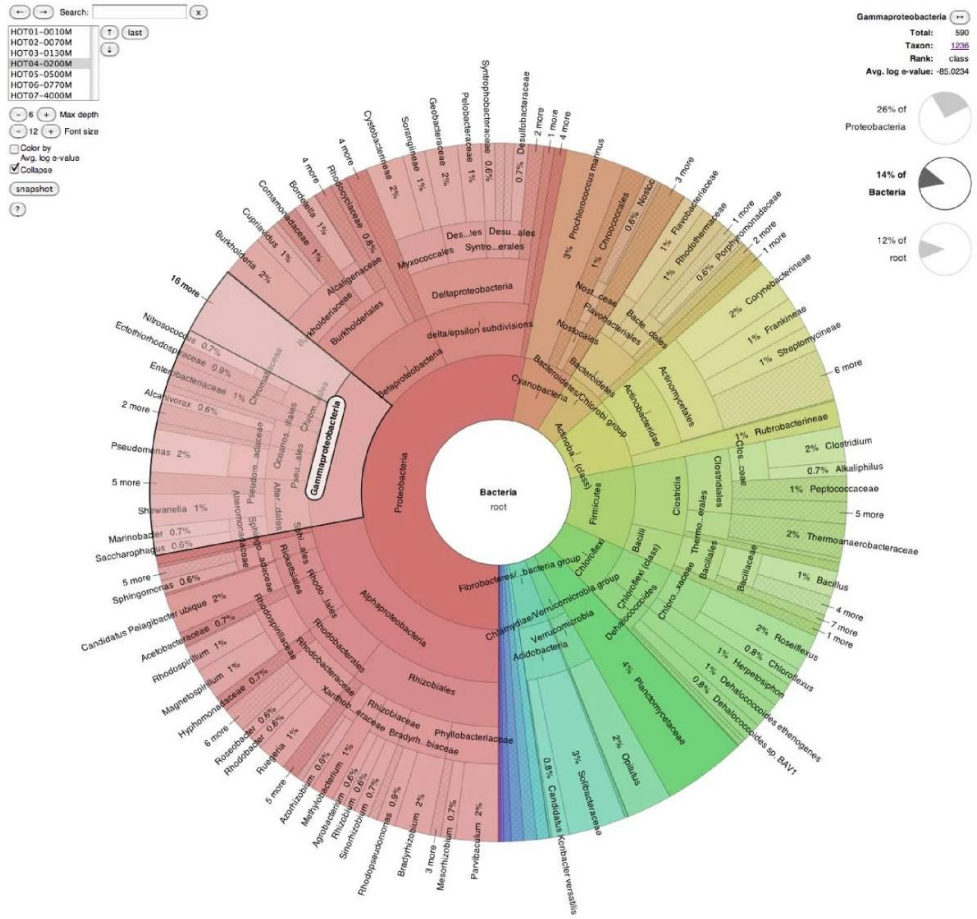
Όσον αφορά την ακρίβεια των εργαλείων, χρησιμοποιήθηκαν οι μετρήσεις μέσου απόλυτου ποσοστού σφάλματος και ανομοιότητας Bray-Curtis. Και οι δύο μπορούν να υπολογίσουν πόσο διαφορετική είναι η προβλεπόμενη διανομή δείγματος σε σχέση με την πραγματική. Ενώ και τα δύο εργαλεία έδωσαν γενικές μετρήσεις reads ανά γένος, το Kraken 2 (Wood et al., 2019) είναι το μόνο εργαλείο το οποίο ευθέως συνδέει κάθε read με την ταξινομική βαθμίδα του μικροοργανισμού στο οποίο αυτό αντιστοιχεί. Το Kraken 2 (Wood et al., 2019) ταξινομεί αναγνώσεις σε κάθε επίπεδο ταξινόμισης, γι' αυτό ορισμένες φορές λανθασμένα θέτει αναγνώσεις σε υψηλότερα επίπεδα γένους, υποτιμά τις αφθονίες των γενών τους. Το QIIME2 (Bolyen et al., 2019) είχε το υψηλότερο ποσοστό σφάλματος κατά την ταξινόμηση του ανθρώπινου δείγματος έναντι του Greengenes (Quast et al., 2013) ή του SILVA (Quast et al., 2013), ανεξάρτητα από το τρόπο μέτρησης ανομοιότητας.

2.2.2 Krona

Για την οπτικοποίηση των αποτελεσμάτων θα χρησιμοποιηθεί τον εργαλείο Krona (Ondov et al., 2011), το οποίο χρησιμοποιεί XML αρχεία για να αποθηκεύσει

πληροφορίες σε μια ιστοσελίδα. Η Krona (Ondov et al., 2011) παρουσιάζει τα γνωρίσματα ως στοιχεία HTML, επιτρέποντας συνδέσμους να συνδέουν σελίδες για κάθε κόμβο και ενσωματώνει στοιχεία javascript ώστε να είναι όσο πιο διαδραστικό γίνεται. Επιπλέον, καθώς τα εργαλεία μεταγονιδιωματικής ανάλυσης συνεχώς βελτιώνονται, το Krona (Ondov et al., 2011) είναι σχεδιασμένο ώστε να είναι ανεξάρτητο αυτών των μεθόδων και να είναι αρκετά ευέλικτο στην ενσωμάτωση νέων.

Το Krona (Ondov et al., 2011) είναι ένα διαδραστικό εργαλείο οπτικοποίησης που επιτρέπει τη διαισθητική εξερεύνηση της σχετικής αφθονίας και της εμπιστοσύνης στις πολύπλοκες ιεραρχίες των μεταγονιδιωματικών ταξινομήσεων. Χρησιμοποιείται σε μορφή αρχείου html το οποίο επιτρέπει την αλληλεπίδραση του χρήστη με τα αποτελέσματα της μεταγονιδιωματικής διεργασίας.



Εικόνα 3. Τυπικό παράδειγμα διαγράμματος Krona (Ondov et al., 2011), παρουσιάζει την ταξινόμηση του Βόρειου Ειρηνικού Υποτροπικού Γύρου εισαγόμενο από το METAREP (Goll et al., 2010). Ο τομέας των βακτηρίων είναι στο επίκεντρο και έχει επιλεχθεί η ταξινομία των ‘Γ-Προτεοβακτηρίων’

2.3 Προϋπάρχουσες Εφαρμογές

2.3.1 NGPhylogeny.fr

Ο κλάδος ενδιαφέροντος της πλατφόρμας ήταν από την αρχή η μεταγονιδιωματική, καθώς είναι σχετικά καινούργιος κλάδος με ποικίλες εφαρμογές. Βασική έμπνευση της εφαρμογής υπήρξε το NGPhylogeny.fr (Lemoine et al., 2019), το οποίο δημιουργήθηκε αρχικά το 2008, σχεδιασμένο για την διευκόλυνση εκτέλεσης φυλογενετικών ροών. Παρ ’όλα αυτά, από την ανάπτυξη του μέχρι τώρα, οι ανάγκες των χρηστών έχουν εξελιχθεί, νέα εργαλεία και ροές έχουν δημοσιευθεί και η ποσότητα των εργασιών έχει αυξηθεί δραματικά, προάγοντας έτσι νέες διεργασίες, το οποίο οδήγησε στην ανακατασκευή του. Το NGPhylogeny.fr (Lemoine et al., 2019) αναπτύχθηκε για να είναι πιο ευέλικτο όσον αφορά τα εργαλεία και τις ροές, εύκολα εγκαταστήσιμο και πιο κλιμακούμενο. Τα διαθέσιμα εργαλεία καλύπτουν μεγάλο εύρος εργασιών (αναζήτηση αλληλουχίας ,στοίχιση πολλαπλών ακολουθιών, επιλογή μοντέλου, συμπερασματικά δέντρα και σχεδίαση δέντρων) και μια μεγάλη ομάδα φυλογενετικών μεθόδων (εξελικτική απόσταση, φειδωλότητα, μέγιστη πιθανοφάνεια και Bayesian). Είναι ενσωματωμένα σε ροές εργασίας, οι οποίες έχουν ήδη διαμορφωθεί («One click»), μπορούν να προσαρμοστούν («Advanced») ή προσαρμόζονται από τον χρήστη («A la carte»). Οι ροές εργασίας διαχειρίζονται και εκτελούνται από ένα υποκείμενο σύστημα ροής εργασιών Galaxy(Afgan et al., 2018), το οποίο κάνει τις ροές εργασίας πιο κλιμακωτές ως προς τον αριθμό των εργασιών και το μέγεθος των δεδομένων.

2.4 Σύνολο δεδομένων Μεταγονιδιωματικής

Για αρχείο εισαγωγής απαιτείται μορφή .fasta (το οποίο μπορεί να βρεθεί και ως .fa, .fas, .ffn, .faa, .frn ή και .fna) ή fastq, στη συγκεκριμένη περίπτωση χρησιμοποιείται ένα toy dataset με κωδικό τρεξίματος SRR8179678 το οποίο προέρχεται από μία κλινική μελέτη παιδιών κάτω της ηλικίας των 18 με αυτισμό και γαστρεντερικές διαταραχές, οι οποίοι υποβλήθηκαν σε θεραπεία μεταμόσχευσης μικροχλωρίδας κοπράνων σε μια προσπάθεια να μειωθεί η σοβαρότητα των συμπτωμάτων συμπεριφοράς και των γαστρεντερικών προβλημάτων τους. Μετρήθηκαν οι διαφορές στο μικροβίωμα τους, μέσω διαφόρων μεταβλητών συμπεριλαμβανόμενου του Parent Global Impressions-III και του Childhood Autism Rating Scale και της σοβαρότητας των γαστρεντερικών συμπτωμάτων τους μέσω της μέτρησης Gastrointestinal Symptom Rating Scale, σε περίοδο 18 εβδομάδων. Τα δείγματα λήφθηκαν από εβδομαδιαία επιχρίσματα κοπράνων, στην ολοκληρωμένη μελέτη η οποία ήταν η πρώτη φάση των κλινικών δοκιμών για τον έλεγχο της ασφάλειας της θεραπείας (Leon et al., 2018).

3. Υλοποίηση Εφαρμογής

3.1 Κώδικας

3.1.1 Front-End

Στην Εικόνα 4 αναπαριστάται το μοντέλο της πλατφόρμας και η διασύνδεση όλων των στοιχείων που έχω αναφέρει έως τώρα. Ξεκινώντας από το στοιχείο άνω αριστερά, οι χρήστες επικοινωνούν μονάχα με το δημόσιο τμήμα (front-end), όπου μπορούν να «ανεβάσουν» τα αρχεία τους για ανάλυση και να επιλέξουν τα εργαλεία ή τις ροές που θα υλοποιήσουν. Πιο συγκεκριμένα, έρχονται σε επαφή με την αρχική σελίδα (homepage), τις σελίδες επιλογής ροών, τις σελίδες των μονών εργαλείων και μια σελίδα επικοινωνίας. Το μοναδικό μέρος του front-end το οποίο δεν είναι διαθέσιμο είναι η σελίδα του διαχειριστή (admin) η οποία είναι εμφανής μόνο σε αυτούς που έχουν πρόσβαση στο συνολικό κώδικα και έχουν ρόλο διαχειριστή.

Όσον αφορά την χρήση της εφαρμογής, η αναλυτική διαδικασία περιγράφεται στην Εικόνα 5, όπου αναφέρεται η επιλογή αρχείου και εισαγωγή αρχείου στο επιλεγμένο εργαλείο ή ροή υπολογισμού. Σε κάθε περίπτωση ζητείται το αρχείο να περιέχει τουλάχιστον 3 αλληλουχίες, με σκοπό την καλή λειτουργία του προγράμματος. Η διαφορά μεταξύ των σελίδων υλοποίησης μονού εργαλείου και ροής είναι πως στην πρώτη περίπτωση υπάρχει μόνο η επεξήγηση της λειτουργίας ενός εργαλείου και οι παράμετροι οι οποίες συνδέονται με αυτό ενώ σε μια ροή περιγράφονται όλα τα βήματα και δίνονται επιλογές για τον χρήστη όπου χρειάζονται. Ο χρήστης καλείται να εισάγει το αρχείο που επιθυμεί να αναλυθεί ακολουθώντας τις συγκεκριμένες προδιαγραφές του εργαλείου και την διεύθυνση ηλεκτρονικού ταχυδρομείου του, σε περίπτωση στην οποία η ολοκλήρωση της διεργασίας που έχει επιλέξει διαρκέσει μεγάλο διάστημα χρόνου να έχει τη δυνατότητα να ενημερωθεί για όταν τελικά ολοκληρωθεί.

Μόλις ο τελικός χρήστης ολοκληρώσει αυτές τις επιλογές, το αρχείο «ανεβαίνει» στην έκδοση (instance) Galaxy (Afgan et al., 2018) που έχουμε επιλέξει. Για να συμβεί αυτό, χρησιμοποιείται το API του Galaxy(Afgan et al., 2018) σε συνδυασμό με το Bioblend (Sloggett et al., 2013), το οποίο είναι φανερό μόνο στον διαχειριστή της εφαρμογής, ο οποίος έχει πρόσβαση στον συνολικό κώδικα (ειδικά το back-end).

3.1.2 Back-End

Από τη πλευρά του back-end, αρχικά αποθηκεύεται προσωρινά τοπικά το αρχείο του χρήστη για λόγους ασφαλείας (μέχρι την ολοκλήρωση της διεργασίας), τοποθετείται σε ένα ειδικό ιστορικό εισόδων/εισαγωγών και δημιουργείται ένα νέο ιστορικό στο Galaxy (Afgan et al., 2018) με μοναδικό αριθμό εντοπισμού όπου θα τρέξουν όλα τα εργαλεία της συγκεκριμένης επίκλησης κάθε ροής /εργαλείου και θα υπάρξουν όλες οι έξοδοι των εργαλείων που χρησιμοποιήθηκαν ώστε ο χρήστης να έχει την επιλογή λήψης και των αρχείων με τα αποτελέσματα των ενδιάμεσων βημάτων μιας διεργασίας.

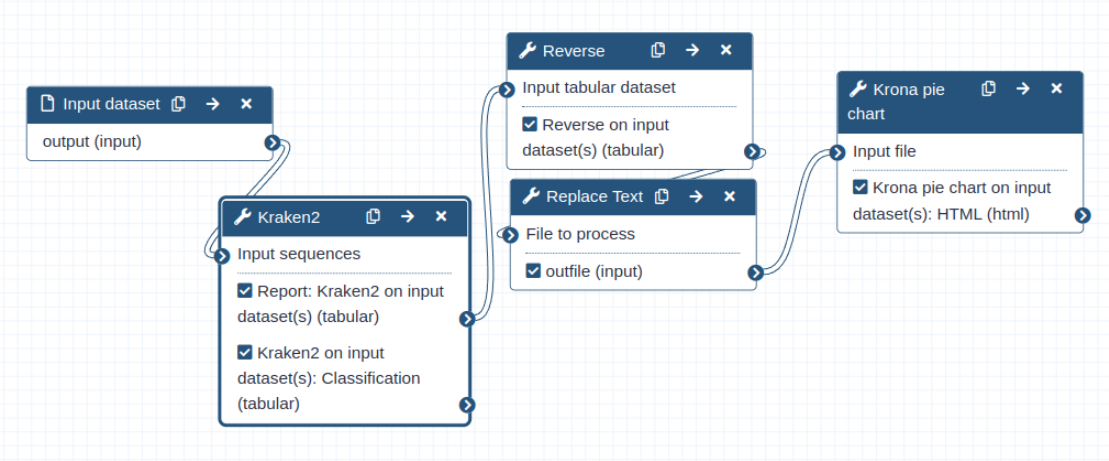
Η διεργασία ξεκινά μόλις αποθηκευτεί το αρχείο στο ιστορικό εισόδων και δημιουργηθεί το ιστορικό για τα αποτελέσματα. Το Galaxy δημιουργεί μια «ουρά υποχρεώσεων» στην οποία κάθε εργαλείο έχει την δική του θέση και περιμένει την

ολοκλήρωση του προηγούμενου για να αρχίσει το επόμενο. Τα εργαλεία που χρησιμοποιούνται είναι προγραμματισμένα από βιοπληροφορικούς διαφόρων ειδικοτήτων, οι οποίοι έχουν προσθέσει το εργαλείο τους στην βιβλιοθήκη conda (βιβλιοθήκη με μεγάλο αριθμό προγραμμάτων, συμπεριλαμβανομένης της βιοπληροφορικής) και έπειτα έχουν γράψει ένα αρχείο json ώστε να προσθέσουν το εργαλείο τους στο ToolShed(Blankenberg et al., 2014) του Galaxy(Afgan et al., 2018). Αυτό επιτρέπει τη χρήση χιλιάδων εργαλείων διαφόρων τομέων της βιοπληροφορικής, συμπεριλαμβανομένης της μεταγονιδιωματικής που αφορά αυτή την εργασία.

Μόλις ολοκληρωθεί η διεργασία όλων των εργαλείων της ουράς δίνεται η επιλογή στον χρήστη να «κατεβάσει» τα αποτελέσματα του, είτε μόνο τα τελικά είτε και τα αποτελέσματα των ενδιάμεσων βημάτων. Ο χρόνος ολοκλήρωσης εξαρτάται από διάφορες παραμέτρους, συγκεκριμένα την πολυπλοκότητα και αριθμό εργαλείων, το μέγεθος του αρχείου εισαγωγής και το πλήθος διεργασιών στην γενική ουρά του Galaxy (Afgan et al., 2018) (διότι χρησιμοποιείται από άλλους χρήστες ταυτόχρονα). Γι' αυτό η εισαγωγή email, υπάρχει περίπτωση, να είναι απαραίτητη για ένα χρήστη με απαιτητική διεργασία.

3.2 Πρακτική Εφαρμογή

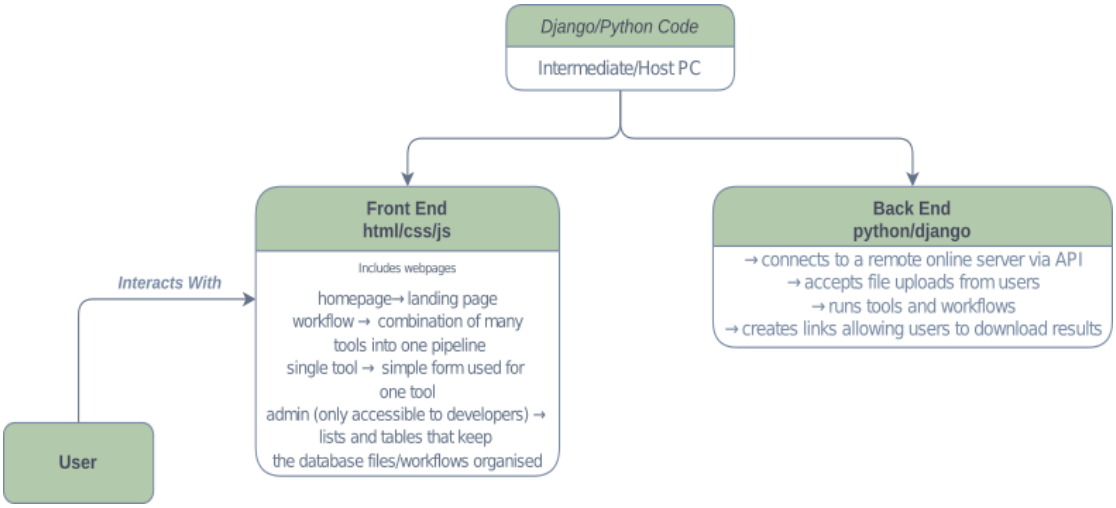
Για καλύτερη επεξήγηση της εφαρμογής χρησιμοποιείται η υλοποίηση μιας ροής εργαλείων η οποία είναι έτοιμη για χρήση, με στόχο την ταξινόμηση μεταγονιδιωματικών αλληλουχιών. Η ροή αυτή περιέχει ως βασικά εργαλεία το Kraken 2 (Wood et al., 2019) και το Krona (Ondov et al., 2011).



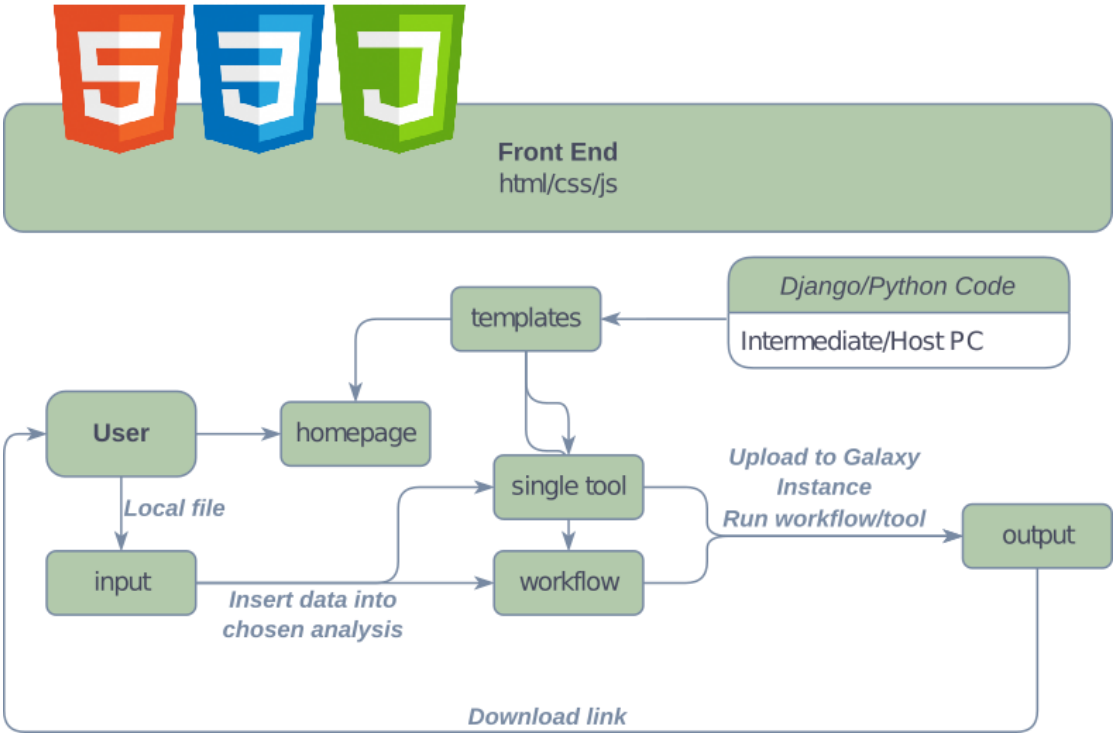
Εικόνα 4. Η ροή εργαλείων που χρησιμοποιείται στα πλαίσια της εφαρμογής. Αρχίζει με εισαγωγή των δεδομένων, ταξινόμησή τους στο Kraken 2 (Wood et al., 2019) με χρήση της βάσης δεδομένων MiniKraken, επεξεργασία των αποτελεσμάτων και παρουσίαση τους σε μορφή εύκολα κατανοητή από το χρήστη και παρουσίαση των αποτελεσμάτων σε πίνακα Krona (Ondov et al., 2011)

Το σύνολο δεδομένων που χρησιμοποιείται είναι μόνο το 1% του συνολικού όγκου δεδομένων και περιλαμβάνει δεδομένα από 5 άτομα τα οποία δέχθηκαν τη θεραπεία και 5 τα οποία ήταν control. Συγκαταλέγονται μεταξύ έξι και δεκαέξι δείγματα από κάθε άτομο, συμπεριλαμβάνοντας τα δείγματα κοπράνων από κάθε άτομο, πριν και μετά της θεραπείας. Τα δεδομένα αλληλουχήθηκαν σε δύο τρεξίματα της Illumina MiSeq.

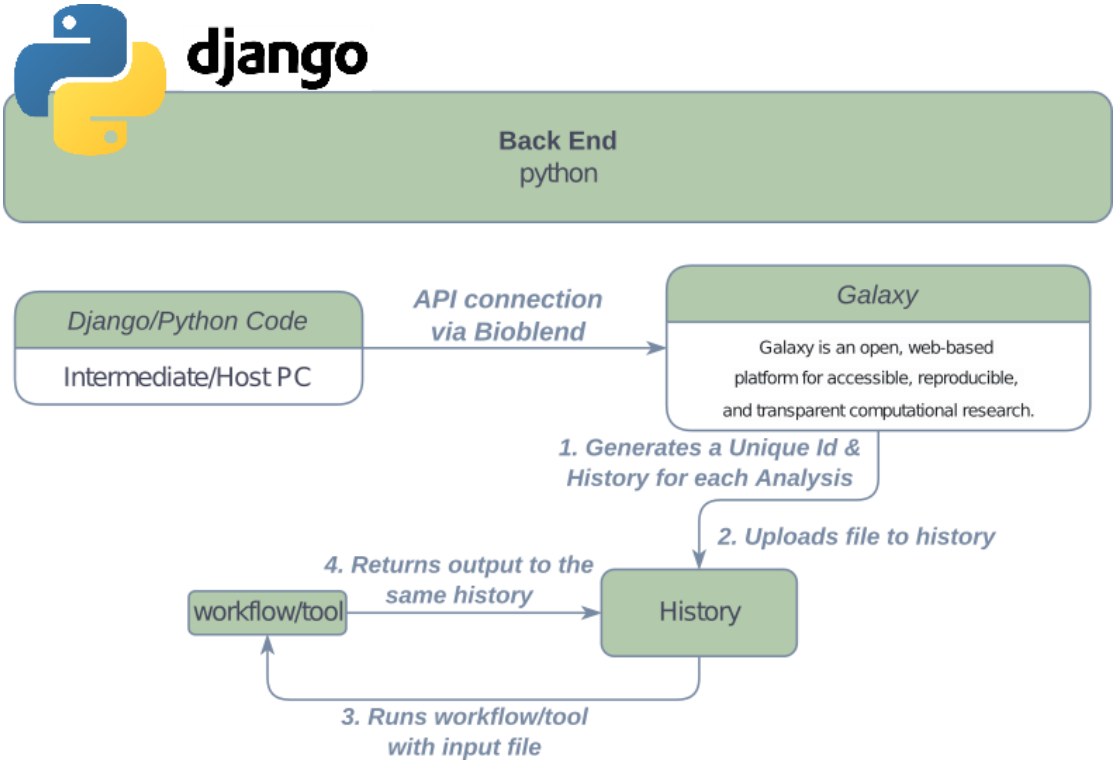
Έπειτα αρχίζουμε την διεργασία της ροής εργαλείων εισάγοντας τα δεδομένα στην αρχική σελίδα. Ως διαχειριστής της εφαρμογής παρατηρούμε πως μόλις πατήσει «υποβολή» (submit) ο χρήστης, εμφανίζεται το αρχείο που υπέβαλε σε έναν ειδικό φάκελο του προγράμματος ως απόδειξη επιτυχής υποβολής και στο σύστημα διαχειριστών εμφανίζεται νέα εγγραφή σε σελίδα με τη φόρμα του συγκεκριμένου εργαλείου που χρησιμοποιήθηκε στη κάθε περίπτωση με αναγνωριστικό το email που δήλωσε ο χρήστης. Όταν πατηθεί η εγγραφή εμφανίζεται πάλι το email και το αρχείο που υπέβαλε. Παράλληλα, αν ελέγξουμε την εκδοχή του Galaxy (Afgan et al., 2018) στην οποία έχουμε συνδεθεί για τη χρήση της ροής εργαλείων, έχει αποθηκευτεί το αρχικό αρχείο σε ένα ιστορικό με όλα τα άλλα αρχικά αρχεία και έχει δημιουργηθεί ένα καινούργιο ιστορικό στο οποίο έχουν ενταχθεί όλα τα βήματα της ροής στα οποία θα αποθηκευτούν τα αποτελέσματα του κάθε εργαλείου που θα τρέξει. Αρχικά αυτά παρουσιάζονται μόνο ως εργαλεία που περιμένουν να τρέξουν διότι έχουν εισαχθεί σε ουρά υλοποίησης. Μόλις ολοκληρωθούν τα εργαλεία, εμφανίζονται τα αποτελέσματα και δίνεται ευκαιρία στον χρήστη να τα κάνουν λήψη.



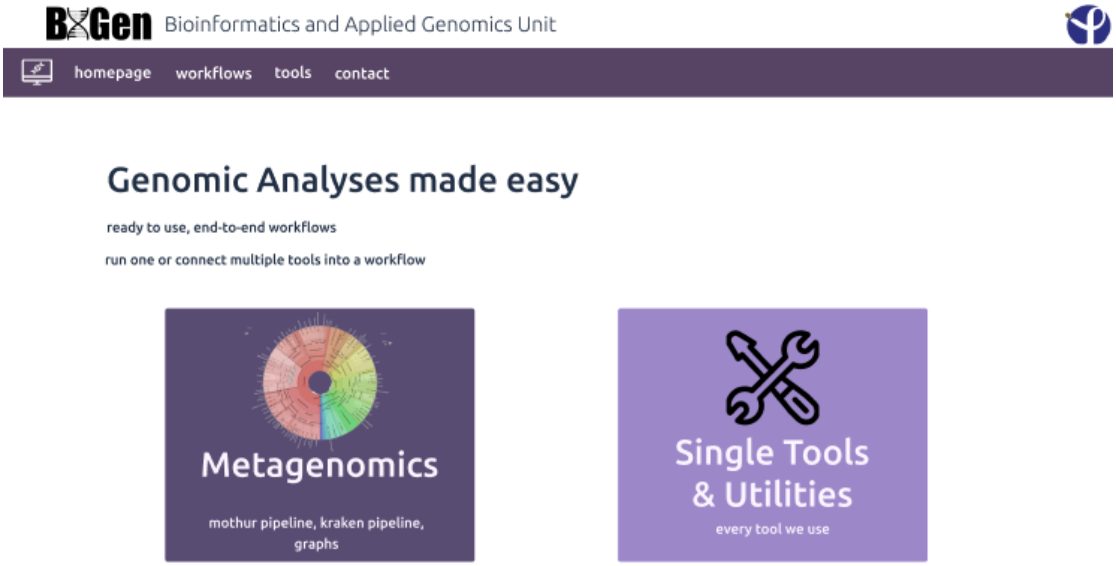
Εικόνα 5. Συνολικό μοντέλο χρήσης της πλατφόρμας συμπεριλαμβάνοντας το front-end και το back-end και το πως συνδέονται μεταξύ τους



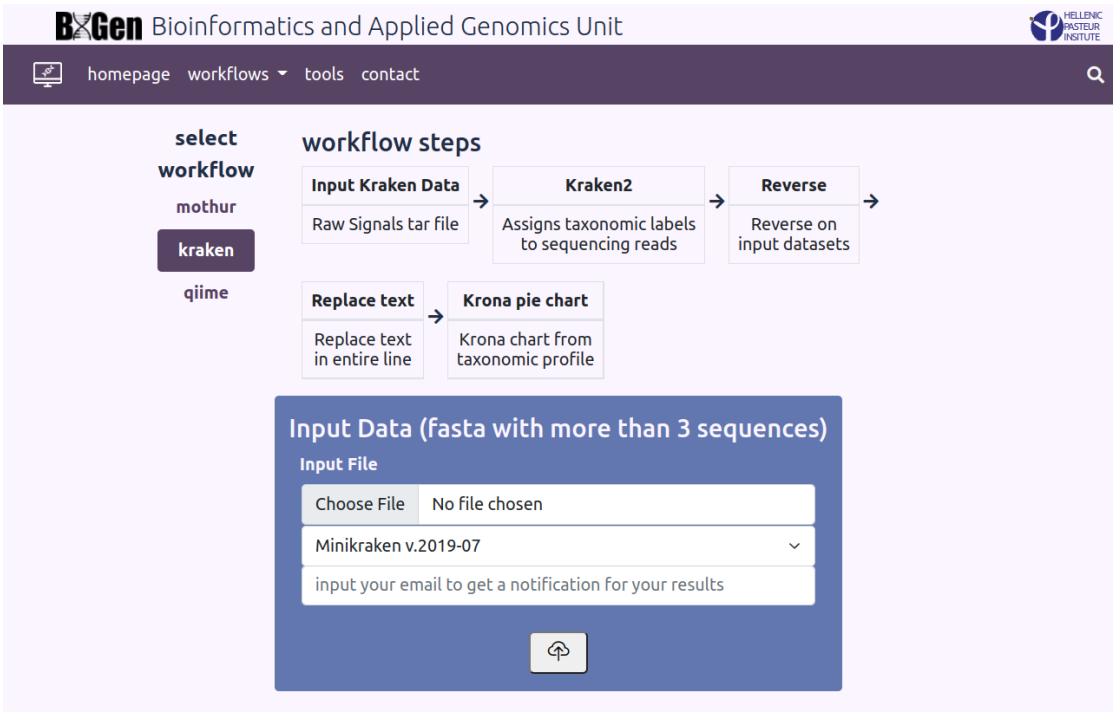
Εικόνα 6. Περιγραφή του front-end, δείχνοντας τις επιλογές που έχουν οι χρήστες στην περιήγηση στην εφαρμογή. Αρχικά, οι χρήστες έχουν μονάχα πρόσβαση στις δημόσιες σελίδες της. Ξεκινώντας από την αρχική σελίδα, έχουν την δυνατότητα να επιλέξουν ανάμεσα στην χρήση ροής εργαλείων και μονού εργαλείου. Έπειτα, ανεβάζουν το αρχείο τους, αναμένουν την ολοκλήρωση της ανάλυσης και τέλος λαμβάνουν τα αποτελέσματα.



Εικόνα 7. Περιγραφή του back-end, αναφέροντας τις σχέσεις που έχουν τα διάφορα σημεία μεταξύ τους από τον αρχικό υπολογιστή όπου αποθηκεύεται ο κώδικας στη σύνδεση του με το Galaxy(Afgan et al., 2018) μέχρι τον χρήστη



Εικόνα 8. Η αρχική σελίδα της εφαρμογής, όπου παρουσιάζονται οι παρόντες διαθέσιμες επιλογές ανάλυσης (μεταγενομική και ξεχωριστά εργαλεία).



Εικόνα 9. Σελίδα τρεξίματος ροής εργαλείων, παρουσιάζεται η επιλογή ροής από τον χρήστη, περιγραφή των βημάτων καθεμιάς και φόρμα υποβολής με πλαίσια για το αρχείο, την βάση δεδομένων που θα χρησιμοποιηθεί και το email του χρήστη.

3.3 Σύγκριση των αποτελεσμάτων που λήφθηκαν από την χρήση της πλατφόρμας με τα αποτελέσματα που λήφθηκαν με χρήση του Terminal

Για να ελεγχθεί η ποιότητα των αποτελεσμάτων της πλατφόρμας σε σχέση με το ίδιο εργαλείο από το terminal, υλοποιήθηκε μια δοκιμή και των δύο μεθόδων με τα ίδια δεδομένα, τα οποία έπειτα συγκρίθηκαν μεταξύ τους. Χρησιμοποιήθηκε επίσης η ίδια βάση δεδομένων (MiniKraken) και λήφθηκαν δύο είδη αρχείων εξόδου, η κανονική έξοδος του προγράμματος που δείχνει την ταξινόμηση και μία έκθεση των αποτελεσμάτων η οποία ταξινομεί τις ταξινομήσεις των οργανισμών με βάσει την πληθώρα τους στο δείγμα.

Η μόνη σημαντική διαφορά που εμφανίστηκε στη σύγκριση της εξόδου αποτελεσμάτων του Kraken 2 (Wood et al., 2019) υπήρξε στη σύνταξη των αρχείων. Στα αποτελέσματα του terminal δινόταν η επιστημονική ονομασία των ανιχνευμένων οργανισμών μαζί με το taxa id τους, σε περίπτωση αγνώστου οργανισμού, δινόταν η περιγραφή ‘unclassified’ με taxa id 0. Στη περίπτωση των αποτελεσμάτων που δόθηκαν από την εφαρμογή, αντί για τις ονομασίες δινόταν μονάχα το taxa id ή 0, στη περίπτωση άγνωστου οργανισμού. Τα υπόλοιπα δεδομένα ταυτίζονταν.

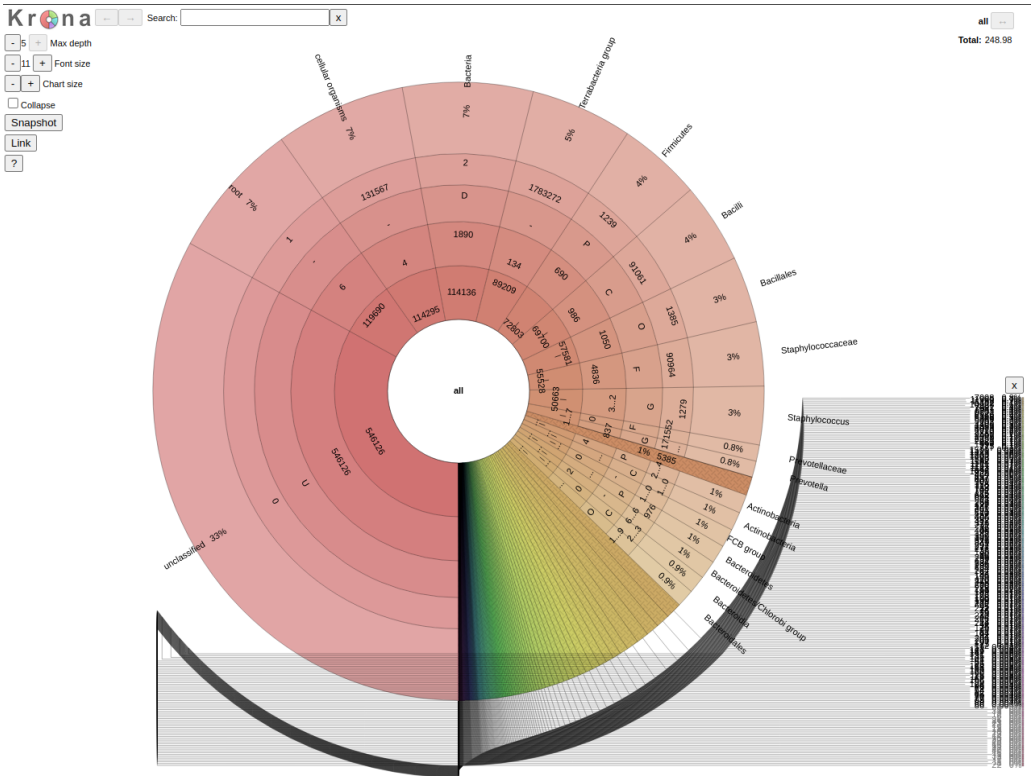
Συγκεκριμένα η δομή των αρχείων εξόδου του Kraken 2 (Wood et al., 2019) είναι δομημένα ως εξής: αρχικά δίνεται μία στήλη η οποία είναι σημασμένη με είτε C είτε U το οποίο δηλώνει το αν είναι ταξινομημένο (classified) ή αταξινόμητο (unclassified). Έπειτα δίνεται η ονομασία του FASTQ/FASTA που αναλύθηκε μαζί με τον αριθμό της αλληλουχίας που εξετάζεται κάθε φορά. Ακολουθεί η στήλη που ήδη περιγράφηκε η οποία έχει διαφορετική μορφή ανάλογα με τον τρόπο ολοκλήρωσης της ανάλυσης. Η έπειτα στήλη δείχνει το μήκος της αλληλουχίας σε βάσεις. Στην περίπτωση αυτής της ανάλυσης υπάρχουν συνδεδεμένα άκρα οπότε δίνεται το μήκος και των δύο χωρισμένα με έναν ειδικό χαρακτήρα, για παράδειγμα ‘126|126’. Τέλος, η τελευταία στήλη δείχνει την χαρτογράφηση των k-mers στην κάθε αλληλουχία.

Για παράδειγμα, στην Εικόνα έχουμε τα αποτελέσματα, συγκεκριμένα από το terminal της αλληλουχίας 21 στην οποία εμφανίζεται το Prevotella intermedia, ένα παθογενές βακτήριο και δίνεται πως τα πρώτα 49 k-mer δεν βρέθηκαν στη βάση δεδομένων όμως τα επόμενα 4 στοιχήθηκαν σε οργανισμό με taxa id 28131. Στη συνέχεια, επιπλέον τέσσερα δεν στοιχήθηκαν, πέντε στοιχήθηκαν πάλι σε οργανισμό με taxa id 28131 και τα τελευταία 30 δε στοιχήθηκαν επιτυχώς. Αφού πρόκειται για αλληλουχίες με συνδεδεμένα άκρα υπάρχει διαχωρισμός μεταξύ τους με τη μορφή αυτών των συμβόλων ‘ |:| ‘ και φαίνεται πως για ακόμα μία φορά υπάρχουν άγνωστες βάσεις σε τρεις περιπτώσεις και αναγνωρίζει ακολουθίες δύο φορές οι οποίες ανήκουν στο οργανισμό με id 28132.

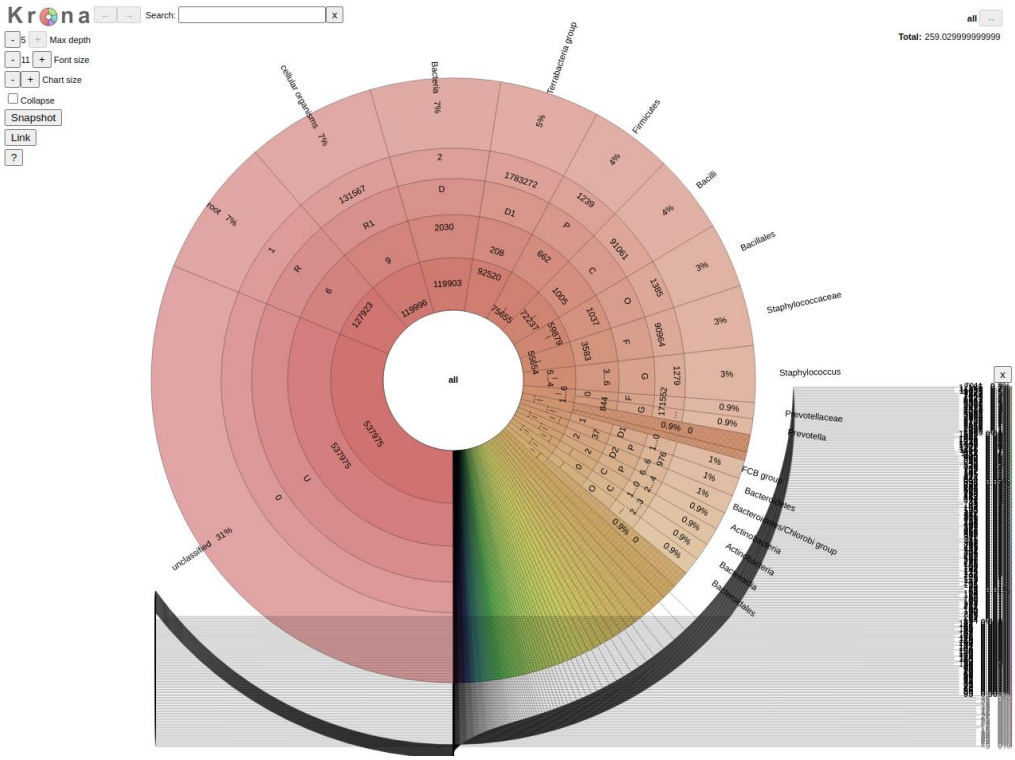
Όσον αφορά την έκθεση αποτελεσμάτων στα αποτελέσματα του terminal και αυτά του εργαλείου υπήρξαν δραστικές διαφορές. Στην περίπτωση του terminal τα αποτελέσματα παρουσιάζονται στις εξής έξι στήλες: πρώτα το ποσοστό θραυσμάτων καλυμμένων από τον κλάδο που έχει τις ρίζες του στη συγκεκριμένη ταξινομική ομάδα, στην δεύτερη στήλη τον αριθμό θραυσμάτων καλυμμένων από τον κλάδο, τρίτον τον αριθμό θραυσμάτων ανατεθειμένων σε αυτή τη ταξινομική ομάδα, τέταρτον ένα κωδικό τάξης ο οποίος είναι κυρίως μεταξύ U(αταξινόμητο), R(έμβια όντα),

D(επικράτεια), K(βασίλειο), P(συνομοταξία), C(τάξη), O(ομοταξία), F(οικογένεια), G(γένος), S(είδος), εάν δεν υπόκεινται σε μια από αυτές τις ομάδες τότε ο κωδικός του διαμορφώνεται από την πιο κοντινή του τάξη μαζί με έναν αριθμό που υποδεικνύει την απόσταση του από αυτή τη τάξη. Πέμπτον υποδεικνύεται ο κωδικός του NCBI και τελικά το επίσημο επιστημονικό του όνομα.

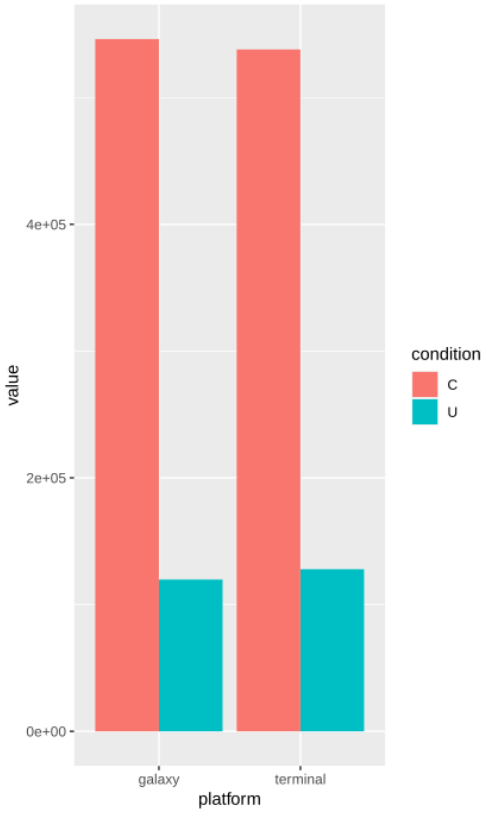
Δεν παρατηρείται κάποια σημαντική διαφορά στα αποτελέσματα των δύο αναλύσεων τόσο σε επίπεδο ταξινόμησης των μικροοργανισμών, όσο και σε επίπεδο χρόνου τρεξίματος.



Εικόνα 10. Αποτελέσματα διαγράμματος Krona (Ondov et al., 2011) με χρήση της εφαρμογής, με συνολικό αριθμό ανιχνευμένων οργανισμών 248.98



Εικόνα 11. Αποτελέσματα διαγράμματος Krona (Ondov et al., 2011) με χρήση του terminal, με συνολικό αριθμό ανιχνευμένων οργανισμών 259.029~



Εικόνα 12. Σύγκριση ανιχνευμένων και μη ανιχνευμένων οργανισμών από τα αποτελέσματα της πλατφόρμας και του terminal

Στην έρευνα που πραγματοποίησαν οι Leon et al. ανίχνευσαν τους ιούς : εντεροϊός A71, ηχοϊός E30,ιός Coxsackie B και ανθρώπινος ερπητοϊός 7. Ο τελευταίος από αυτούς βρέθηκε μονάχα με χρήση τεχνολογιών mNGS, ενώ οι υπόλοιποι ανιχνεύθηκαν με PCR και mNGS. Όμως υπήρχαν περιπτώσεις στις οποίες οι ιοί ανιχνεύθηκαν μονάχα με mNGS. Πραγματοποιήθηκε σύγκριση των αποτελεσμάτων μεταξύ της έρευνας, του terminal και της πλατφόρμας στα δείγματα εγκεφαλονωτιαίου υγρού και τα αποτελέσματα επιβεβαίωσαν τα ευρήματα της αρχικής μελέτης. Βρέθηκε πως τα αποτελέσματα της ανάλυσης μέσω Kraken 2 στο terminal αλλά και στην πλατφόρμα, ανίχνευσαν τους ίδιους ιούς, συγκεκριμένα τον εντεροϊό A71, τον ιό Coxsackie B και τον ανθρώπινο ερπητοϊό 7. Αυτό δείχνει πως η ανάλυση δεδομένων μπορεί να πραγματοποιηθεί στα ίδια επίπεδα επιτυχίας και από τις δύο μεθόδους.

Viruses	Analysis Methods		
	Platform	Terminal	PCR-confirmed study from Leon et al., 2018
Enterovirus A71 (EV-A71)	Βρέθηκε	Βρέθηκε	Βρέθηκε
Echovirus 30 (E30)	Δεν Βρέθηκε	Δεν Βρέθηκε	Βρέθηκε
Coxsackie Virus B (CVB)	Βρέθηκε	Βρέθηκε	Βρέθηκε
Human betaherpesvirus 7 (HHV-7)	Βρέθηκε	Βρέθηκε	Βρέθηκε

Πίνακας 2. Ανίχνευση ιών στα δείγματα εγκεφαλονωτιαίου υγρού από την έρευνα (Leon et al., 2018) μέσω των αποτελεσμάτων της έρευνας και ανάλυσης του Kraken2 στο terminal και στην πλατφόρμα

4. Συμπεράσματα

Η ανάπτυξη της μεταγονιδιωματικής και πιο συγκεκριμένα της κλινικής μεταγονιδιωματικής έχει το ενδεχόμενο να βελτιώσει τεχνικές διάγνωσης και θεραπείας. Για να εκπληρωθούν πλήρως αυτές οι προοπτικές απαιτούνται βελτιώσεις στην τεχνολογία αλληλούχισης και σε βιοπληροφορικά εργαλεία που σχετίζονται με την ταξινόμηση των αλληλουχιών σε μικροοργανισμούς. Μέσω της πλατφόρμας που έχει αναπτυχθεί παρέχεται η δυνατότητα αναβαθμισμένης γενωμικής ανάλυσης τόσο σε έμπειρους χρήστες όσο και σε ερευνητές διαφόρων ειδικοτήτων (π.χ. διαφόρων ειδικοτήτων Ιατροί, Βιολόγοι, Βιοπληροφορικοί) οι οποίοι δεν διαθέτουν τις απαιτούμενες γνώσεις βιοπληροφορικής. Ενώ η χρήση της τώρα διατίθεται μόνο σε τοπικό επίπεδο, έχει τεθεί ως απώτερος στόχος η δημοσίευσή της δημόσια στο διαδίκτυο και η ενσωμάτωση νέων εργαλείων που εξυπηρετούν άλλους κλάδους της επιστήμης της βιολογίας.

Η ολοκλήρωση της εργασίας δεν υπήρξε χωρίς προκλήσεις και δυσκολίες, που αφορούν τόσο το μέρος του προγραμματισμού όσο και τη διαχείριση των εργαλείων και των ροών.

Αρχικά αυτή υπήρξε η πρώτη εργασία σε γλώσσα Python που ολοκλήρωσα και στην ουσία όλη η εκμάθησή μου σε αυτή τη γλώσσα υπήρξε μέσω της διεκπεραίωσης αυτής της εργασίας. Ενώ είναι γνωστή ως γενικώς «εύκολη» γλώσσα, αυτό δε σημαίνει πως δεν υπάρχουν εμπόδια στην πλήρης χρήση της. Μπορώ να πω ειλικρινά πως ακόμα μαθαίνω την βέλτιστη χρήση της για διαφορετικές εφαρμογές. Η Python έχει πολλαπλές χρήσεις και θα ήθελα να συνεχίσω να την χρησιμοποιώ στο μέλλον.

Βασικό πρόβλημα εμφανίστηκε με τη χρήση λογισμικού ανεπτυγμένου από άλλους προγραμματιστές. Συγκεκριμένα, καθώς η πλατφόρμα που ανέπτυξα χρησιμοποιεί διεπαφή προγραμματισμού εφαρμογών (API) για την βασική λειτουργία της, αναγκάστηκα να βασιστώ σε κώδικα που δεν είχα γράψει προσωπικά, τον οποίο δεν μπορούσα να αλλάξω με σκοπό να διευκολύνω την εφαρμογή μου. Επιπλέον υπήρξαν καταστάσεις στις οποίες ο κώδικας του API άλλαζε και αναγκαζόμουν να αλλάξω κρίσιμα σημεία του δικού μου και να αναζητήσω νέους τρόπους λύσεις των προβλημάτων. Το ίδιο ίσχυε και με τις ενημερώσεις της Python και Django οι οποίες θα δημιουργούσαν σφάλματα τα οποία θα χρειάζονταν χρόνο για την επίλυσή τους κατά τη διάρκεια της ανάπτυξης.

Αναφορικά με τα εργαλεία, η χρήση πολλών έχει επιπλέον εξαρτήσεις, συνήθως αυτό αναφέρεται σε βάσεις δεδομένων οι οποίες αν δεν υπάρχουν ήδη εγκατεστημένες στην εκδοχή που χρησιμοποιείται κάνουν τη χρήση του εργαλείου αδύνατη. Γι’αυτό το λόγο η επιλογή της κατάλληλης εκδοχής είναι υψίστης σημασίας.

Βιβλιογραφία

Django Software Foundation, 2019. Django, Available at: https://djangoproject.com. https://doi.org/10.1038/nmeth.4458

Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379

Applications, N.R.C. (US) C. on M.C. and F., 2007. Why Metagenomics?, The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. National Academies Press (US).

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., Taylor, J., Nekrutenko, A., Galaxy Team, 2014. Dissemination of scientific software with Galaxy ToolShed. Genome Biol. 15, 403. https://doi.org/10.1186/gb4161

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., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., vander Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Chiu, C.Y., Miller, S.A., 2019. Clinical metagenomics. Nat. Rev. Genet. 20, 341–355. https://doi.org/10.1038/s41576-019-0113-7

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, D633–D642. https://doi.org/10.1093/nar/gkt1244

d'Humières, C., Salmona, M., Dellière, S., Leo, S., Rodriguez, C., Angebault, C.,

Alanio, A., Fourati, S., Lazarevic, V., Woerther, P.L., Schrenzel, J., Ruppé, E., 2021. The Potential Role of Clinical Metagenomics in Infectious Diseases: Therapeutic Perspectives. Drugs 81, 1453–1466. https://doi.org/10.1007/s40265-021-01572-4

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072. https://doi.org/10.1128/AEM.03006-05

Goll, J., Rusch, D.B., Tanenbaum, D.M., Thiagarajan, M., Li, K., Methé, B.A., Yooseph, S., 2010. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. Bioinformatics 26, 2631–2632. https://doi.org/10.1093/bioinformatics/btq455

Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L., 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729. https://doi.org/10.1101/gr.210641.116

Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G., Knight, R., 2011. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al CHAPTER, Unit10.7. https://doi.org/10.1002/0471250953.bi1007s36

Leo, S., Pireddu, L., Cuccuru, G., Lianas, L., Soranzo, N., Afgan, E., Zanetti, G., 2014. BioBlend.objects: metacomputing with Galaxy. Bioinformatics 30, 2816–2817. https://doi.org/10.1093/bioinformatics/btu386

Leon, K.E., Casas-Alba, D., Ramesh, A., Khan, L.M., Launes, C., Sample, H.A., Zorn, K.C., Valero-Rello, A., Langelier, C., Muñoz-Almagro, C., DeRisi, J.L., Wilson, M.R., 2018. Pediatric Brainstem Encephalitis Outbreak Investigation with Metagenomic Next-Generation Sequencing. https://doi.org/10.1101/414979

Levy, S.E., Myers, R.M., 2016. Advancements in Next-Generation Sequencing. Annu. Rev. Genomics Hum. Genet. 17, 95–115. https://doi.org/10.1146/annurev-genom-083115-022413

Loeffler, C., Gibson, K.M., Martin, L., Chang, L., Rotman, J., Toma, I.V., Mason, C.E., Eskin, E., Zackular, J.P., Crandall, K.A., Koslicki, D., Mangul, S., 2019. Unlocking the diagnostic and therapeutic potential of metagenomics 1–59.

López-Labrador, F.X., Brown, J.R., Fischer, N., Harvala, H., Van Boheemen, S., Cinek, O., Sayiner, A., Madsen, T.V., Auvinen, E., Kufner, V., Huber, M., Rodriguez, C., Jonges, M., Hönemann, M., Susi, P., Sousa, H., Klapper, P.E., Pérez-Cataluña, A., Hernandez, M., Molenkamp, R., der Hoek, L. van, Schuurman, R., Couto, N., Leuzinger, K., Simmonds, P., Beer, M., Höper, D., Kamminga, S., Feltkamp, M.C.W., Rodríguez-Díaz, J., Keyaerts, E., Nielsen, X.C., Puchhammer-Stöckl, E., Kroes, A.C.M., Buesa, J., Breuer, J., Claas, E.C.J., de Vries, J.J.C., 2021. Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. J. Clin. Virol. 134. https://doi.org/10.1016/j.jcv.2020.104691

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., Schäffer, A.A., 2008. Database indexing for production MegaBLAST searches. Bioinforma. Oxf. Engl. 24, 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J.,

Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733-745. https://doi.org/10.1093/nar/gkv1189

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385. https://doi.org/10.1186/1471-2105-12-385

Ounit, R., Wanamaker, S., Close, T.J., Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16. https://doi.org/10.1186/s12864-015-1419-2

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 Res. 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., Demaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiutė, M., Hansen, L.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.H., Liao, Y.C., Silva, G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., McHardy, A.C., 2017. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nat. Methods 14, 1063–1071. https://doi.org/10.1038/nmeth.4458

Sloggett, C., Goonasekera, N., Afgan, E., 2013. BioBlend: Automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29, 1685–1686. https://doi.org/10.1093/bioinformatics/btt199

Walker, M.A., Pedamallu, C.S., Ojesina, A.I., Bullman, S., Sharpe, T., Whelan, C.W., Meyerson, M., 2018. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics 34, 4287–4289. https://doi.org/10.1093/bioinformatics/bty501

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Ye, S.H., Siddle, K.J., Park, D.J., Sabeti, P.C., 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178, 779–794. https://doi.org/10.1016/j.cell.2019.07.010