



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Ανάλυση δεδομένων μεταγονιδιωματικής (metagenomics)
και μικροβιώματος (microbiome) από πειράματα NGS με την
τεχνική του quasi-mapping**

Γιώργος Ν. Σκούφος

Επιβλέπουσα: **Καθ. Άρτεμις Χατζηγεωργίου**, Ερευνήτρια, Τμήμα Μηχανικών
Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου
Θεσσαλίας, Επικεφαλής του DIANA-Lab.

ΑΘΗΝΑ

ΙΑΝΟΥΑΡΙΟΣ 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση δεδομένων μεταγονιδιωμικής (metagenomics) και μικροβιώματος (microbiome) από πειράματα NGS με την τεχνική του quasi-mapping

Γιώργος Ν. Σκούφος

A.M.: ΠΙΒ0145

ΕΠΙΒΛΕΠΟΥΣΑ: **Καθ. Άρτεμις Χατζηγεωργίου**, Ερευνήτρια, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας, Επικεφαλής του DIANA-Lab.

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Καθ. Άρτεμις Χατζηγεωργίου**, Ερευνήτρια, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας, Επικεφαλής του DIANA-Lab.
Δρ. Ιωάννης Βλάχος, Επιστημονικός Συνεργάτης, Broad Institute του MIT και Harvard, Brigham and Women's Hospital.
Δρ. Μαρία Παρασκευοπούλου, Ερευνήτρια/Επιστημονικός Συνεργάτης, DIANA-Lab, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.

Ιανουάριος 2017

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη των τεχνολογιών αλληλούχισης επόμενης γενιάς (Next Generation Sequencing - NGS) έχουν μετατρέψει την ικανότητα μας για διερεύνηση της σύνθεσης και δυναμικής των μικροβιακών κοινοτήτων που κατοικούν στα χερσαία και υδάτινα οικοσυστήματα, καθώς επίσης και στο ανθρώπινο δέρμα, το έντερο και το στόμα. Τα δεδομένα μεταγονιδιωματικής (metagenomics) που προκύπτουν από πειράματα NGS, συνήθως περιλαμβάνουν ένα μεγάλο αριθμό από μικροοργανισμούς (βακτήρια, ιούς κ.τ.λ.) και ως εκ τούτου, συνήθως, παράγουν αρχεία πολύ μεγάλου μεγέθους.

Σκοπός της παρούσας εργασίας ήταν ο σχεδιασμός και η ανάπτυξη ενός υπολογιστικού εργαλείου το οποίο να έχει τη δυνατότητα να εντοπίζει και να ποσοτικοποιεί τους οργανισμούς σε επίπεδο υποειδών (subspecies, strains) σε σύνθετα δείγματα μεταγονιδιωματικής και μικροβιώματος (microbiome) από πειράματα NGS. Το εργαλείο έχει τη δυνατότητα να χρησιμοποιηθεί σε δείγματα που προέρχονται από πειράματα όπως τα 16S rRNA Sequencing και Shotgun Metagenomic Sequencing καθώς επίσης και τη δυνατότητα του εντοπισμού και ποσοτικοποίησής του μικροβιώματος σε μικτά δείγματα ιστού (tissue-specific) DNA/RNA όπου αποτελούνται από τον ξενιστή (άνθρωπος, ποντίκι, άλλα θηλαστικά είδη) και το μικροβίωμα του.

Οι βασικές λειτουργίες του εργαλείου που αναπτύχθηκε είναι ο εντοπισμός και ποσοτικοποίηση του μικροβιώματος σε δείγματα NGS, ο υπολογισμός της αφθονίας των μικροβίων στις ταξινομικές βαθμίδες (taxonomic ranks) της οικογένειας, του γένους, του είδους και των υποειδών και τέλος το φιλτράρισμα των αποτελεσμάτων με κριτήρια που ορίζονται από τον χρήστη.

Η εργασία παρουσιάζει τα αποτελέσματα που προέκυψαν από το εργαλείο σε συνθετικά αλλά και πραγματικά δεδομένα. Από την σύγκριση που έγινε με άλλα εργαλεία μεταγονιδιωματικής, φαίνεται, πως, σε όλες τις περιπτώσεις, το εργαλείο που αναπτύχθηκε παράγει πιο ακριβή αποτελέσματα και σε πολλές περιπτώσεις είναι ταχύτερο.

Το metaHost είναι ένα γρήγορο και με μεγάλη ακρίβεια εργαλείο το οποίο εντοπίζει και ποσοτικοποιεί μικροβιακούς οργανισμούς σε σύνθετα NGS δείγματα μεταγονιδιωματικής, με πλήρη αυτόματο τρόπο και με μεγάλη προσαρμοστικότητα στις ανάγκες κάθε χρήστη.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μεταγονιδιωματική, μικροβίωμα, αλληλούχιση επόμενης γενιάς, NGS, quasi-mapping, εντοπισμός μικροβίων, ποσοτικοποίηση

ABSTRACT

The development of high-throughput sequencing technologies has transformed our capacity to investigate the composition and dynamics of the microbial communities that populate terrestrial and aquatic ecosystems as well as the human skin, gut and oral cavity. Sequenced metagenomic samples usually comprise reads from a large number of different bacterial and viral communities, and hence tend to result in vast file sizes.

The purpose of the present study was the design and implementation of a computational tool - pipeline which has the ability to identify and quantify organisms at strain level in complex Microbiomic, Metagenomic, and Metatranscriptomic Next Generation Sequencing (NGS) samples. The pipeline has the ability to be used as a metagenome classifier in 16S rRNA Sequencing and Shotgun Metagenomic Sequencing datasets as well as the ability to analyze mixed tissue-specific DNA/RNA NGS samples consisting of the host (Human, Mouse, other mammalian species) and its microbiome.

The main functions of the implemented pipeline are the identification and quantification of the microbiome in NGS samples, the abundance estimation in Family, Genus, Species and Subspecies taxonomic ranks and the filtering of the estimated results based on user-specific criteria.

This study presents the results obtained by applying the pipeline to analyze both microbiome and mixed host-microbiome simulated NGS datasets as well as real tissue-specific *Mus Musculus* RNA datasets obtained from NCBI's GEO. The comparison between the implemented pipeline and state-of-the-art metagenomic classification tools, showed, that in every case, the pipeline produces more accurate results in terms of abundance estimation and in many cases, is faster too.

metaHost is a rapid and accurate pipeline that identifies and quantifies microbiome organisms at strain level in complex Metagenomic – Metatranscriptomic NGS samples based on a fully automated workflow which is easily adaptable to the needs of its users.

SUBJECT AREA: Bioinformatics

KEYWORDS: metagenomics, microbiome, metatranscriptomics, Next Generation Sequencing, reads mapping, NGS, microbiome identification, quantification, quasi-mapping

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα της διπλωματικής μου εργασίας, καθηγήτρια του Πανεπιστημίου Θεσσαλίας κυρία Χατζηγεωργίου για την ευκαιρία που μου έδωσε να ασχοληθώ με το παρόν θέμα. Ευχαριστώ ιδιαίτερα τον ερευνητή της Ιατρικής Σχολής του Χάρβαρντ, κύριο Ιωάννη Βλάχο για τις συμβουλές, την υπομονή και την καθοδήγησή του, χάρη στα οποία κατάφερα να ολοκληρώσω τη συγγραφή της παρούσας διπλωματικής εργασίας. Ευχαριστώ θερμά τη μεταδιδακτορικό του Πανεπιστημίου Θεσσαλίας, Μαρία Παρασκευοπούλου, για τις υποδείξεις της σε σχέση με το πρόγραμμα που αναπτύχθηκε. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την στήριξη που μου παρέχει όλα αυτά τα χρόνια και τη Σοφία.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ	12
1.1 Το μικροβίωμα	12
1.2 Το ανθρώπινο μικροβίωμα.....	13
1.3 Βακτήρια	14
1.4 Ιοί.....	15
1.5 Αρχαία.....	16
1.6 Μεταγονιδιωμική.....	17
1.7 Το πρόβλημα.....	18
1.8 Τεχνολογίες μελέτης μικροβιώματος.....	19
1.8.1 Καλλιέργεια μικροβίων	19
1.8.2 Αλυσιδωτή αντίδραση πολυμεράσης (PCR)	19
1.8.3 Ταξινόμηση βακτηρίων με χρήση του 16S ριβοσωματικού RNA	19
1.8.4 Απευθείας αλληλούχιση του 16S rRNA (16S rRNA-Seq)	20
1.8.5 Μαζική αλληλούχιση μεταγονιδιώματος (Shotgun metagenomic sequencing).....	20
1.9 State of the art.....	22
1.9.1 MetaPhlAn2.....	22
1.9.2 Kraken.....	23
1.9.3 Clark.....	24
1.9.4 Bracken	24
1.9.5 Kallisto.....	24
1.10 Σκοπός.....	26
2. ΜΕΘΟΔΟΙ.....	28
2.1 Το Pipeline.....	28
2.2 Αρχείο παραμέτρων (configuration file)	28
2.3 Έλεγχος αρχείων εισόδου.....	28
2.4 Στοίχιση των δειγμάτων στο γονιδίωμα / μεταγράψωμα του ξενιστή	29
2.5 Απομάκρυνση των ακολουθιών του ξενιστή από το δείγμα	29

2.6	Απομάκρυνση του PhiX 174 από το δείγμα	29
2.7	Εντοπισμός του μικροβιώματος	30
2.8	Υπολογισμός αφθονίας στις ταξινομικές βαθμίδες του γένους, είδους και υποείδους	31
2.9	Φιλτράρισμα των αποτελεσμάτων	32
2.10	Υπολογισμός βασικών στατιστικών	32
2.11	Οπτικοποίηση αποτελεσμάτων.....	32
2.12	Αρχεία εξόδου.....	32
3.	ΣΥΝΘΕΤΙΚΑ ΔΕΔΟΜΕΝΑ.....	34
3.1	Δημιουργία συνθετικών δεδομένων	34
3.2	Αποτελέσματα του metaHost	36
3.3	Σύγκριση του metaHost – Γραφήματα.....	50
3.4	Σύγκριση του metaHost – Μετρικές (metrics).....	60
3.5	Χρόνοι εκτέλεσης.....	66
4.	ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ.....	67
4.1	Τα δεδομένα	67
4.2	Η βάση των μικροβίων (reference).....	68
4.3	Αποτελέσματα.....	68
5.	ΣΥΜΠΕΡΑΣΜΑΤΑ	77
5.1	Συμπεράσματα	77
	ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ.....	79
	ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ.....	82
	ΑΝΑΦΟΡΕΣ.....	83

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1 - metaHost Dataset 1 - Genus level	37
Σχήμα 2 - metaHost Dataset 1 - Species level	38
Σχήμα 3 - metaHost Dataset 1 - Strain level.....	39
Σχήμα 4 - metaHost Dataset 2 - Genus level	40
Σχήμα 5 - metaHost - Dataset 2 - Species level	41
Σχήμα 6 - metaHost Dataset 3 - Genus level	42
Σχήμα 7 - metaHost Dataset 3 - Species level	43
Σχήμα 8 - metaHost Dataset 4 - Genus level	44
Σχήμα 9 - metaHost Dataset 4 - Species level	45
Σχήμα 10 - metaHost Dataset 4 - Strain level.....	46
Σχήμα 11 - metaHost Dataset 5 - Genus level	47
Σχήμα 12 - metaHost Dataset 5 - Species level	48
Σχήμα 13 - metaHost Dataset 5 - Strain level.....	49
Σχήμα 14 - Σύγκριση αλγόριθμων Dataset 1 - Genus level	50
Σχήμα 15 - Σύγκριση αλγόριθμων Dataset 1 - Species level.....	51
Σχήμα 16 - Σύγκριση αλγόριθμων Dataset 2 - Genus level	52
Σχήμα 17 - Σύγκριση αλγόριθμων Dataset 2 - Species level.....	53
Σχήμα 18 - Σύγκριση αλγόριθμων Dataset 3 - Genus level	54
Σχήμα 19 - Σύγκριση αλγόριθμων Dataset 3 - Species level.....	55
Σχήμα 20 - Σύγκριση αλγόριθμων Dataset 4 - Genus level.....	56
Σχήμα 21 - Σύγκριση αλγόριθμων Dataset 4 - Species level.....	57
Σχήμα 22 - Σύγκριση αλγόριθμων Dataset 5 - Genus level.....	58
Σχήμα 23 - Σύγκριση αλγόριθμων Dataset 5 - Species level.....	59

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1 – Δομή ενός τυπικού Βακτηρίου	14
Εικόνα 2 - Δομή ενός τυπικού Ιού.....	15
Εικόνα 3 - Δομή ενός τυπικού Αρχαίου.....	16
Εικόνα 4 - Τα βήματα της τεχνικής Shotgun Metagenomic Sequencing	21
Εικόνα 5 - Ο αλγόριθμος του Kraken.....	23
Εικόνα 6 - Η τεχνική του pseudo-alignment (kallisto).....	25
Εικόνα 7 - Η τεχνική του quasi-mapping	31
Εικόνα 8 - Το Pipeline σχηματικά	33

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1 - Τα συνθετικά δεδομένα που δημιουργήθηκαν	35
Πίνακας 2 - Metrics dataset 1 - Genus level	61
Πίνακας 3 - Metrics dataset 1 - Species level	61
Πίνακας 4 - Metrics dataset 2 - Genus level	62
Πίνακας 5 - Metrics dataset 2 - Species level	62
Πίνακας 6 - Metrics dataset 3 - Genus level	63
Πίνακας 7 - Metrics dataset 3 - Species level	63
Πίνακας 8 - Metrics dataset 4 - Genus level	64
Πίνακας 9 - Metrics dataset 4 - Species level	64
Πίνακας 10 - Metrics dataset 5 - Genus level	65
Πίνακας 11 - Metrics dataset 5 - Species level	65
Πίνακας 12 - Χρόνοι εκτέλεσης αλγόριθμων	66
Πίνακας 13 - Πληροφορίες των δεδομένων που αναλύθηκαν	67
Πίνακας 14 - Αποτελέσματα dataset PRJNA246715	69
Πίνακας 15 - Αποτελέσματα dataset PRJNA279201	69
Πίνακας 16 - Αποτελέσματα dataset PRJNA193402	70
Πίνακας 17 - Αποτελέσματα dataset PRJNA169535	70
Πίνακας 18 - Αποτελέσματα dataset PRJNA248400	71
Πίνακας 19 - Αποτελέσματα dataset PRJNA257651	71
Πίνακας 20 - Αποτελέσματα dataset PRJNA254957	71
Πίνακας 21 - Αποτελέσματα dataset PRJNA177791	72
Πίνακας 22 - Αποτελέσματα dataset GSM900198 rep1.....	72
Πίνακας 23 - Αποτελέσματα dataset GSM900198 - rep2	73
Πίνακας 24 - Αποτελέσματα dataset GSM900189 rep2.....	73
Πίνακας 25 - Αποτελέσματα dataset GSM900186 rep1.....	74

Πίνακας 26 - Αποτελέσματα dataset GSM900186 rep2.....	74
Πίνακας 27 - Αποτελέσματα dataset PRJNA285523 (16S rRNA).....	75
Πίνακας 28 - Μικρόβια που εντοπίστηκαν σε ιστοειδικά και μικροβιακά δείγματα	76

1. ΕΙΣΑΓΩΓΗ

1.1 Το μικροβίωμα

Μικροοργανισμοί χαρακτηρίζονται εκείνοι οι οργανισμοί τους οποίους δεν μπορούμε να διακρίνουμε με γυμνό μάτι, γιατί έχουν μέγεθος μικρότερο από 0,1mm [1]. Ο χώρος που ασχολείται με τους μικροοργανισμούς ονομάζεται μικροβιολογία ενώ οι πρώτοι μικροοργανισμοί ανακαλύφθηκαν το 1674 από τον Antonie van Leeuwenhoek χρησιμοποιώντας ένα μικροσκόπιο δικής του κατασκευής. Οι μικροοργανισμοί απαντώνται σε κάθε περιοχή της βιόσφαιρας, όπως στο έδαφος, σε θερμές πηγές, στον ωκεανό, στην ατμόσφαιρα, στο εσωτερικό βράχων, και κάτω από τον εσωτερικό φλοιό της γης [2]. Τέλος, υπό ορισμένες συνθήκες και με βάση δοκιμές που έχουν γίνει, έχει παρατηρηθεί πως μικροοργανισμοί επιβιώνουν και στο κενό του διαστήματος [3].

Το μικροβίωμα (microbiome) αποτελείται από όλα τα μικροβιακά κύτταρα που κατοικούν σε μια συγκεκριμένη περιοχή [4]. Όλα τα φυτά και τα ζώα, από τα πρώτιστα (protists) έως τον άνθρωπο ζουν σε στενή συνεργασία με μικροβιακούς οργανισμούς. Μέχρι σχετικά πρόσφατα, οι βιολόγοι είχαν καθορίσει τις αλληλεπιδράσεις των φυτών και των ζώων με τον μικροβιακό κόσμο μόνο στο πλαίσιο συσχετίσεων με ασθένειες και αυτό μονάχα σε ένα πολύ μικρό αριθμό περιπτώσεων συμβίωσης του ξενιστή με το μικροβίωμα του.

Τα τελευταία χρόνια, ένας σημαντικός αριθμός εξελίξεων έχει αλλάξει εντελώς την αντίληψη μας για τους μικροβιακούς οργανισμούς και το ρόλο τους.

Μεταξύ των εξελίξεων αυτών, περιλαμβάνονται οι εξής:

- Η δυνατότητα εκτέλεσης γονιδιωμικών (genomics) αναλύσεων σε ολόκληρες μικροβιακές κοινότητες στο νέο κλάδο της μεταγονιδιωμικής (metagenomics).
- Μαζικές βάσεις δεδομένων που καθιστούν τις σχετικές πληροφορίες προσβάσιμες από ερευνητές πολλών κλάδων.
- Αναλύσεις βιοπληροφορικής που βοηθούν στην εξαγωγή συμπερασμάτων από σύνθετα σύνολα δεδομένων (datasets).

Μετά τις εξελίξεις των τελευταίων ετών, όλο και περισσότεροι εκτιμούν ότι τα μικρόβια διαδραματίζουν σημαντικό ρόλο στο φαινότυπο ενός πολυκύτταρου οργανισμού [5].

1.2 Το ανθρώπινο μικροβίωμα

Στην περίπτωση του ανθρώπου, το μικροβίωμα περιλαμβάνει περισσότερα από 100 τρισεκατομμύρια βακτήρια, ιούς, αρχαία και μύκητες οι οποίοι ζουν εντός ή επί του σώματος [6]. Ουσιαστικά ο αριθμός των μικροβίων που κατοικούν στο ανθρώπινο σώμα, είναι περίπου δέκα φορές μεγαλύτερος από τον συνολικό αριθμό των ανθρωπίνων κυττάρων.

Η μικροχλωρίδα (microbiota) κάθε ανθρώπου είναι μοναδική, παρόλα αυτά, οι φυσιολογικές λειτουργίες της, είναι συνήθως ίδιες και έχουν μεγάλο αντίκτυπο στην υγεία του.

Μερικές από τις λειτουργίες αυτές είναι οι εξής:

- Βοηθάει το σώμα στην πέψη ορισμένων τροφών που το στομάχι και το λεπτό έντερο δεν είναι σε θέση να αφομοιώσουν.
- Βοηθάει στην παραγωγή των βιταμινών B και K.
- Βοηθάει στην καταπολέμηση επιθέσεων από άλλους επιβλαβείς μικροοργανισμούς διατηρώντας την ακεραιότητα του εντερικού βλεννογόνου.
- Διαδραματίζει πάρα πολύ σημαντικό ρόλο στο ανοσοποιητικό σύστημα.

Παρόλο που ο άνθρωπος φέρει μικρόβια σχεδόν σε όλο του το σώμα, υπάρχουν κάποια όργανα στα οποία περιέχονται πολύ περισσότερα σε σχέση με τα υπόλοιπα. Τα όργανα αυτά, είναι συνήθως εκείνα που έρχονται σε επαφή με το εξωτερικό περιβάλλον.

Έτσι, τα περισσότερα μικρόβια στον άνθρωπο, βρίσκονται στη στοματική κοιλότητα, στο δέρμα, στο παχύ και λεπτό έντερο, στους πνεύμονες και στο γυναικείο κόλπο. Φυσικά σε κάθε μια από τις παραπάνω περιοχές, τα είδη των μικροοργανισμών που βρίσκονται σε μεγάλη αφθονία διαφέρουν. Γενικά, το μεγαλύτερο μέρος των μικροοργανισμών που περιέχονται στο ανθρώπινο σώμα αποτελείται από βακτήρια και ακολουθούν σε πολύ μικρότερα ποσοστά οι ιοί, οι μύκητες και τα αρχαία. Οι πληροφορίες αυτές, συνεχώς αλλάζουν και προσαρμόζονται καθώς είναι πολλές οι μελέτες που ασχολούνται με την χαρτογράφηση του ανθρωπίνου μικροβιώματος.

1.3 Βακτήρια

Τα βακτηριόφυτα, ευρύτερα γνωστά ως βακτήρια, είναι μικροσκοπικοί, μονοκύτταροι, προκαρυωτικοί οργανισμοί, που συναντιούνται σε κάθε είδους βióτοπο και σε πολύ μεγάλους αριθμούς, όπως στο χώμα, στον ωκεανό, εντός και επί του ανθρωπίνου σώματος Κ.Ο.Κ. (κώδικας οδικής κυκλοφορίας;)

Τα βακτήρια αποτελούνται κατά μονάδα από ένα κύτταρο και μπορεί να εμφανίζονται με διάφορα σχήματα όπως:

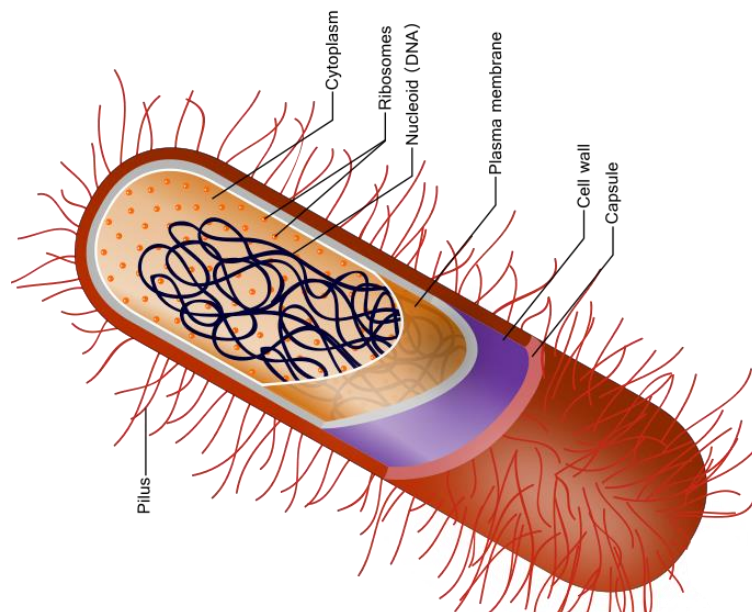
- Σφαιρικό (χαρακτηρίζονται ως κόκκοι).
- Ραβδοειδές (χαρακτηρίζονται ως βάκιλλοι).
- Σπειροειδές (χαρακτηρίζονται ως σπειρίλια).
- Καμπυλόγραμμες ράβδοι (χαρακτηρίζονται ως δονάκια).

Οι διαστάσεις των βακτηρίων μετριοούνται σε μικρόμετρα, (μm), κατά συνέπεια είναι ορατά μόνο με μικροσκόπιο. Πρόκειται για τους μικρότερους μονοκύτταρους οργανισμούς μεταξύ των έμβιων όντων της Φύσης, μετά από τους ιούς και τις συγγενικές ρικέτσιες. (είναι έμβια όντα οι ιοί;)

Τα βακτήρια διακρίνονται σε δύο βασικές κατηγορίες: στα «θετικά κατά Γκραμ» (Gram-positive) και στα «αρνητικά κατά Γκραμ» (Gram-negative). Στην ομάδα των θετικών κατά Γκραμ περιλαμβάνονται οι ακτινομύκητες, οι βάκιλλοι, οι γαλακτοβάκιλλοι, οι στρεπτόκοκκοι και τα κλωστρίδια.

Τα βακτήρια έχουν πολύ απλή κυτταρική δομή. Η κυτταρική τους μεμβράνη αποτελείται από ημικυτταρίνες ή πηκτινοειδείς ουσίες, συνεπώς δεν υφίσταται κυτταρικός πυρήνας καθώς και σύνθετα οργανίδια. Η κυτταρική μεμβράνη των βακτηρίων περιβάλλεται από το κυτταρικό τοίχωμα, που είναι το κοινό χαρακτηριστικό σχεδόν όλων των βακτηρίων.

Το DNA τους οργανώνεται σε πυρηνοειδές, χωρίς πυρηνική μεμβράνη και χωρίς σχηματισμό χρωμοσωμάτων. Πολλά βακτήρια περιέχουν επιπλέον πλασμίδια, δηλαδή μικρά, αυτόνομα κυκλικά μόρια DNA. Η ικανότητα των πλασμιδίων να μεταφέρονται από το ένα βακτήριο στο άλλο έχει τεράστια σημασία στην ανάπτυξη παθογονικότητας και ανθεκτικότητας σε αντιβιοτικά από βακτηριακά στελέχη.



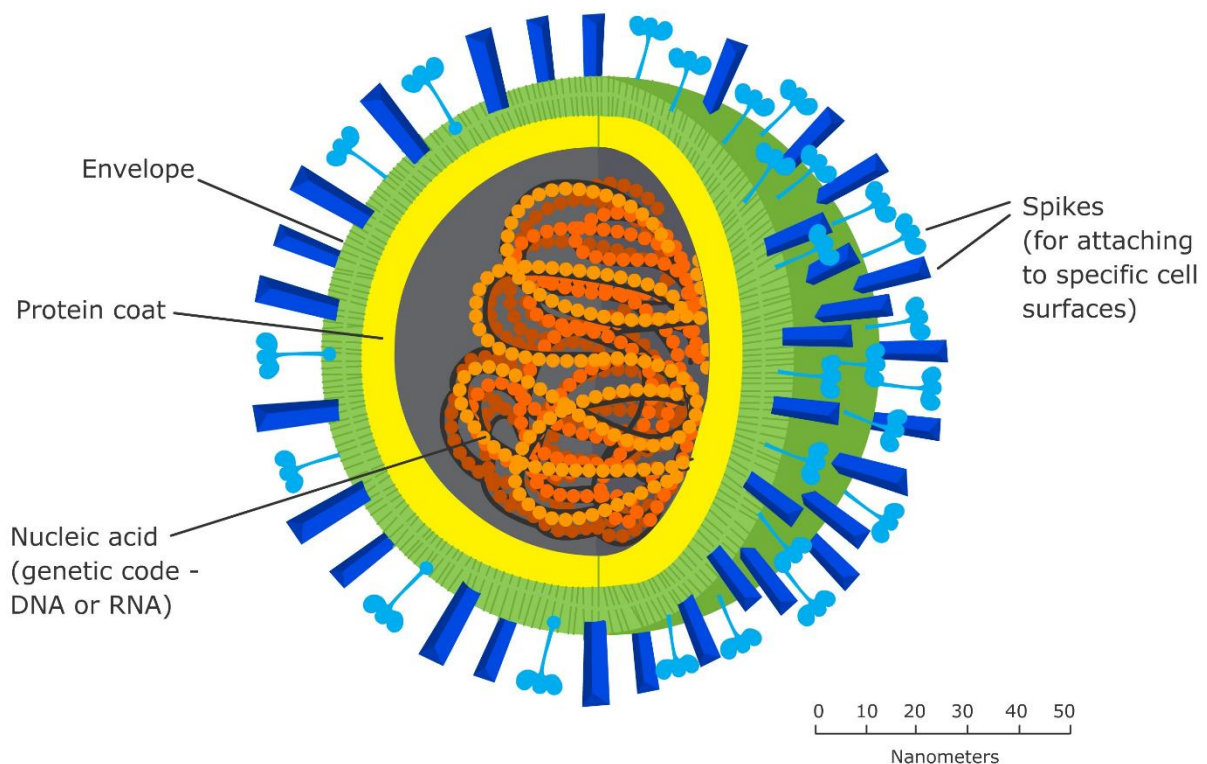
Εικόνα 1 – Δομή ενός τυπικού Βακτηρίου

1.4 Ιοί

Ο ιός δρα μολύνοντας τα κύτταρα ενός οργανισμού, ενσωματώνοντας το γενετικό του υλικό στο γονιδίωμα αυτών και χρησιμοποιώντας τους μηχανισμούς του γονιδιώματος του ξενιστή για τον πολλαπλασιασμό του, τους μηχανισμούς αντιγραφής, μεταγραφής και μετάφρασης του κυττάρου, όπως και τα περισσότερα ένζυμα που χρειάζεται για την επιβίωσή του.

Οι ιοί χαρακτηρίζονται, κατά συνέπεια, ως ενδοκυτταρικά παράσιτα των οποίων το μέγεθος κυμαίνεται από 0,025 μm μέχρι 0,25 μm . Η ύπαρξή τους διαπιστώθηκε για πρώτη φορά τον 19ο αιώνα και παρατηρήθηκαν στο ηλεκτρονικό μικροσκόπιο κατά τον 20ό αιώνα, όταν έγινε δυνατή η κρυστάλλωσή τους.

Κατά γενική αποδοχή οι ιοί μεταλλάσσονται και εξελίσσονται, αν και δεν υπάρχει γενική συμφωνία για το αν μπορούν να θεωρηθούν έμβια συστήματα [7]. Η μελέτη του γενετικού υλικού διαφόρων ιών έδειξε περισσότερες ομοιότητες με το γενετικό υλικό των ξενιστών τους παρά με το γενετικό υλικό άλλων ιών.



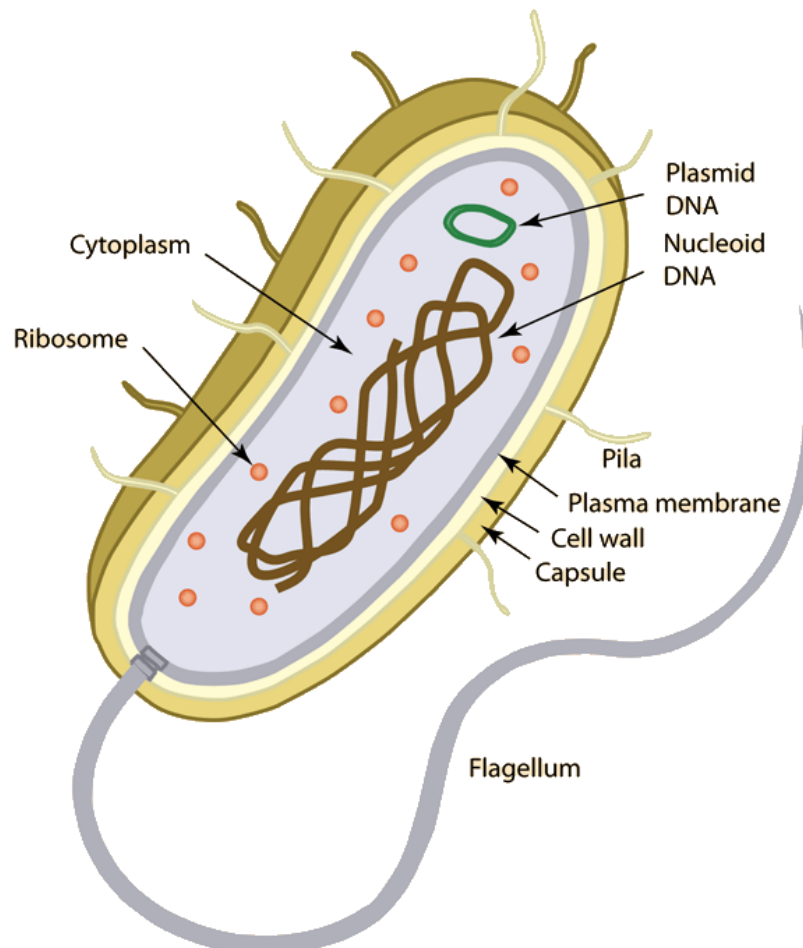
Εικόνα 2 - Δομή ενός τυπικού Ιού

1.5 Αρχαία

Στη Μικροβιολογία αρχαία κατέληξε να χαρακτηρίζονται τα πρώην αρχαιοβακτήρια, καθώς έχουν διαφορετική εξελικτική ιστορία από τους υπόλοιπους μονοκύτταρους μικροοργανισμούς.

Τα αρχαία αποτελούν μια από τις τρεις βασικές ομάδες τελευταίας ταξινόμησης των οργανισμών που έχει γίνει με βάση γενετικές δομές και αλληλουχίες. Συγκεκριμένα αποτελούν τη μεσαία τάξη από τα Βακτήρια και τα Ευκάρυα.

Αν και μοιάζουν πολύ με τα βακτήρια, τα αρχαία διαφέρουν από τα βακτήρια κυρίως στη δομή των λιπιδίων της μεμβράνης τους, και στο μεταβολισμό τους, στη μεταγραφή και τη μετάφραση του RNA, καθώς επίσης και στα κυτταρικά τοιχώματα. Για πηγή ενέργειας τα αρχαία εκμεταλλεύονται μία πολλή μεγάλη ποικιλία στοιχείων, από οργανικές ενώσεις όπως η ζάχαρη μέχρι αμμωνία, ιόντα μετάλλων και υδρογόνο. Στο ανθρώπινο σώμα, τα συναντάμε σε πολύ μικρότερο βαθμό σε σχέση με τα βακτήρια και τους ιούς. Επίσης, ο αριθμός των γονιδιωμάτων των αρχαίων που έχουμε στη διάθεση μας, είναι πολύ μικρότερος από τα δύο παραπάνω.



Εικόνα 3 - Δομή ενός τυπικού Αρχαίου

1.6 Μεταγονιδιωματική

Ο όρος «Μεταγονιδιωματική» (metagenomics) χρησιμοποιήθηκε πρώτη φορά από τους Jo Handelsman, Jon Clardy, Robert M. Goodman και Sean F. Brady και εμφανίστηκε για πρώτη φορά σε μια δημοσίευση του 1998 [8]. Ο όρος μεταγονιδίωμα (metagenome), αναφέρεται στην ιδέα ότι μια συλλογή γονιδίων, τα οποία έχουν υποστεί αλληλούχιση (Sequencing) απευθείας από περιβαλλοντικό δείγμα (environmental sample), μπορούν να αναλυθούν με ανάλογο τρόπο όπως όταν γίνεται ανάλυση ενός ενιαίου γονιδιώματος (genome).

Γενικά, η Μεταγονιδιωματική, είναι η μελέτη του γενετικού υλικού που έχει ανακτηθεί απευθείας από περιβαλλοντικά δείγματα. Στην παραδοσιακή μικροβιολογία (microbiology), την αλληλούχιση μικροβιακών γονιδιωμάτων και την γονιδιωματική (genomics), οι ερευνητές βασίζονται συνήθως στην καλλιέργεια μικροβιακών κοινοτήτων (microbial culture) με σκοπό την μελέτη των οργανισμών. Οι πρόσφατες εξελίξεις στο χώρο της μεταγονιδιωματικής, επιτρέπουν την στοχευμένη αλληλούχιση συγκεκριμένων γονιδίων όπως του 16S rRNA με σκοπό την παραγωγή του προφίλ της ποικιλομορφίας (profile of diversity) ενός φυσικού δείγματος. Με αυτή την εξέλιξη, οι γνώσεις μας για το πώς συνθέτονται οι μικροβιακές κοινότητες έχουν αναπτυχθεί σε πολύ μεγάλο βαθμό [9] συγκριτικά με τα προηγούμενα χρόνια.

Εκτός της τεχνικής της στοχευμένης αλληλούχισης συγκεκριμένων γονιδίων, μια ακόμα τεχνική αλληλούχισης που έχει συμβάλει πολύ στο χώρο της μεταγονιδιωματικής είναι η τυχαία αλληλούχιση του μεταγονιδιώματος (Shotgun metagenomic Sequencing). Στην συγκεκριμένη τεχνική, σε πρώτη φάση γίνεται τυχαία η διάτμηση του DNA σε μικρά μέρη, στην συνέχεια γίνεται μαζικά η αλληλούχιση των μερών αυτών και τέλος από τα μικρά κομμάτια προκύπτουν μεγαλύτερες ακολουθίες.

Εκτενής αναφορά για τις δύο προαναφερθέντες τεχνικές θα γίνει σε επόμενο κεφάλαιο της διπλωματικής εργασίας.

1.7 Το πρόβλημα

Το πρόβλημα το οποίο προσπάθησε να λύσει η παρούσα διπλωματική εργασία, είναι ο εντοπισμός (identification) και η ποσοτικοποίηση (abundance estimation) του DNA/RNA μικροβιακών οργανισμών σε επίπεδο υποειδών (sub-species, strains) σε σύνθετα δείγματα μεταγονιδιωματικής (metagenomic samples) και μικροβιώματος (microbiome) που προέρχονται από πειράματα αλληλούχισης επόμενης γενιάς (next generation sequencing) αλλά και σε μικτά δείγματα που απαρτίζονται από τον ξενιστή (host) όπου μπορεί να είναι ο άνθρωπος, το ποντίκι ή άλλος θηλαστικός οργανισμός, μαζί με το μικρόβιο που φέρει.

Μια από τις βασικές προκλήσεις στο χώρο της μεταγονιδιωματικής και συγκεκριμένα στον εντοπισμό των μικροοργανισμών σε δείγματα αλληλούχισης επόμενης γενιάς, προέρχεται από το γεγονός πως έχουμε στην κατοχή μας δεκάδες χιλιάδες μικροβιακά γονιδιώματα και συνεπώς το ερώτημα που προκύπτει είναι το ποια γονιδιώματα θα χρησιμοποιήσουμε για να δημιουργήσουμε τη βάση (reference) έναντι της οποίας θα γίνει η διαδικασία της στοίχισης των ακολουθιών (alignment).

Το Εθνικό Κέντρο Βιοτεχνολογίας της Αμερικής (National Center for Biotechnology Information - NCBI) περιέχει στην βάση δεδομένων του τις ακολουθίες 3.788 ευκαρυωτικών οργανισμών, 85.156 προκαρυωτικών οργανισμών, 6.179 ιών και τέλος 7.755 πλασμιδίων. Από τις συνολικά 102.878 ακολουθίες, οι 21.427 είναι ολοκληρωμένες (completed genomes) και οι υπόλοιπες 81.451 είναι στο στάδιο επεξεργασίας και διόρθωσής (draft genomes). Τα περισσότερα από τα προαναφερθέντα γονιδιώματα, συγκαταλέγονται σε μικροβιακούς οργανισμούς και συνεπώς έχουν μεγάλο ενδιαφέρον στο χώρο της μεταγονιδιωματικής. Έτσι, ανάλογα με τα δεδομένα που αναλύονται κάθε φορά και του τί μας ενδιαφέρει να ερευνήσουμε σε αυτά, θα πρέπει να δημιουργείται και η κατάλληλη βάση προσαρμοσμένη στις ανάγκες του εκάστοτε ερευνητή.

Η επόμενη και ίσως μεγαλύτερη πρόκληση είναι το γεγονός πως πολλοί από τους μικροοργανισμούς και κυρίως τα βακτήρια, έχουν υψηλή ομοιότητα (high similarity) στις ακολουθίες των γονιδιωμάτων τους [9]. Δεδομένου και του ποσοστού λάθους που έχουμε στις τεχνολογίες NGS αλλά και λόγω των μεταλλάξεων που μπορεί να υπάρχουν στα γονιδιώματα των βακτηρίων, γίνεται αντιληπτό πως ένας αλγόριθμος μπορεί εύκολα να υποπέσει σε λάθος και να αποφανθεί πως ένα read προέρχεται από ένα γονιδίωμα αλλά στην πραγματικότητα, αυτό το read, να προέρχεται από άλλο γονιδίωμα, υψηλής ομοιότητας με το προηγούμενο.

Μια δυσκολία που συναντάμε συχνά στην ανάλυση δεδομένων μεταγονιδιωματικής είναι πως τα δείγματα που αναλύουμε είναι πολύ μεγάλου μεγέθους. Πρακτικά, αυτό σημαίνει, πως χρειάζεται καλή υποδομή υπολογιστικών πόρων (resources) και τεράστιοι χώροι αποθήκευσης (Data storage) για τα δεδομένα αλλά και για τα αποτελέσματα που προκύπτουν από αυτά.

Τέλος, λόγω του μεγάλου μεγέθους των δεδομένων και λόγω της πολυπλοκότητας αυτών, το πρόβλημα έπρεπε να αντιμετωπιστεί με ένα αλγόριθμο υψηλής ακρίβειας και παράλληλα υψηλής ταχύτητας ειδικά η ανάλυση ενός συνόλου δεδομένων θα έπαιρνε πάρα πολλές υπολογιστικές ώρες.

Το υπολογιστικό εργαλείο που προέκυψε από την παρούσα διπλωματική εργασία, επιλύει ικανοποιητικά τα παραπάνω προβλήματα όπως αυτό θα διαπιστωθεί από τα αποτελέσματα που προέκυψαν αναλύοντας συνθετικά αλλά και πραγματικά δεδομένα.

1.8 Τεχνολογίες μελέτης μικροβιώματος

1.8.1 Καλλιέργεια μικροβίων

Για πάρα πολλά χρόνια, οι γνώσεις μας για τα μικρόβια προέρχονταν μόνο από τις τεχνικές καλλιέργειας και απομόνωσης μικροβίων, τεχνικές οι οποίες βοήθησαν πολύ στα πρώτα στάδια μελέτης των μικροοργανισμών. Η καλλιέργεια μικροβίων αποτελεί μια χαμηλού κόστους αλλά παράλληλα χρονοβόρα διαδικασία για την μελέτη μικροβιακών κοινοτήτων. Οι δυνατότητες της εν λόγω τεχνικής είναι αρκετά περιορισμένες καθώς ακόμα και σήμερα, το ποσοστό των μικροβίων που έχει καλλιεργηθεί επιτυχώς σε εργαστηριακούς χώρους, υπολογίζεται περίπου στο 20% του συνόλου των οργανισμών που γνωρίζουμε. Οι εξελίξεις των τελευταίων ετών στο χώρο της καλλιέργειας όπως για παράδειγμα οι "έξυπνοι" επωαστήρες (Smart incubators) αντιμετωπίζουν ικανοποιητικά αρκετά από τα προβλήματα που υπήρχαν στο παρελθόν και φαίνεται πως η απόδοση της συγκεκριμένης τεχνικής θα φέρει περισσότερα αποτελέσματα στο μέλλον [10].

1.8.2 Αλυσιδωτή αντίδραση πολυμεράσης (PCR)

Η γνωστή τεχνική της αλυσιδωτής αντίδρασης πολυμεράσης, είναι μια τεχνική που πλέον χρησιμοποιείται και στην μελέτη του μικροβιώματος. Ανιχνεύοντας στοχευμένα γονίδια μικροοργανισμών, είναι δυνατή η μελέτη μικροχλωρίδων. Ένα από τα βασικά μειονεκτήματα της εν λόγω τεχνικής, είναι η μειωμένη ακρίβεια που έχει στην ποσοτικοποίηση του γενετικού υλικού (DNA) του κάθε οργανισμού στο δείγμα και άρα ο υπολογισμός του πληθυσμού κάθε μικροβίου καθίσταται δύσκολος. Το μειονέκτημα αυτό, αντιμετωπίστηκε από την αλυσιδωτή αντίδραση πραγματικού χρόνου (Real time PCR). Μερικά από τα βασικά πλεονεκτήματα της rPCR είναι η ταχύτητα, η ικανότητα φυλογενετικού διαχωρισμού και τέλος ο σχεδιασμός εκκινήτων που στοχεύουν συγκεκριμένα είδη μικροβίων.

1.8.3 Ταξινόμηση βακτηρίων με χρήση του 16S ριβοσωματικού RNA

Η υπομονάδα 16S του ριβοσωματικού RNA (16S rRNA), χρησιμοποιείται για την μελέτη και ταξινόμηση των βακτηρίων. Το 16S rRNA είναι πολύ καλά συντηρημένο (conserved) και εντοπίζεται σε όλα τα βακτήρια που γνωρίζουμε μέχρι τώρα. Το μήκος του ανέρχεται σε περίπου 1500 νουκλεοτίδια και η βασική του λειτουργία είναι η έναρξη και επιμήκυνση της πρωτεϊνοσύνθεσης. Η λειτουργία της πρωτεϊνοσύνθεσης δεν έχει αλλάξει και συνεπώς οι αλλαγές που έχει υποστεί η ακολουθία του 16S rRNA αποτελούν δείκτη του χρόνου εξέλιξης των βακτηρίων. Το 16S rRNA έχει υψηλή ομολογία μεταξύ των βακτηριακών ειδών, όμως, στην ακολουθία του περιλαμβάνονται εννέα περιοχές οι οποίες μεταβάλλονται ανάλογα με το είδος του βακτηρίου και άρα μπορούν να χρησιμοποιηθούν για αναγνώριση και διάκριση μεταξύ των διαφόρων ειδών. Προτού έρθουν οι τεχνολογίες της μεταγονιδιωμικής, η ταυτοποίηση των μικροοργανισμών με την συγκεκριμένη μέθοδο, γινόταν με πολλαπλασιασμό του 16S rRNA με PCR, κλωνοποίηση και τέλος καθορισμό της αλληλουχίας με την μέθοδο κατά Sanger.

Στη μελέτη των μικροβιακών κοινοτήτων, έχουν συμβάλει και οι εξής τεχνικές:

- Τεχνική αποτύπωσης (fingerprinting)

- Τεχνικές υβριδοποίησης ανιχνευτή (probe hybridization techniques)
- FISH
- Τεχνική μικροσυστοιχιών DNA (DNA microarrays).

1.8.4 Απευθείας αλληλούχιση του 16S rRNA (16S rRNA-Seq)

Καθώς οι τεχνολογίες αλληλούχισης επόμενης γενιάς εξελίσσονται, οι εφαρμογές τους έχουν επεκταθεί και στην έρευνα μικροβιακών κοινοτήτων. Μια από τις τεχνικές αλληλούχισης που έφερε επανάσταση στο χώρο, ήταν η αλληλούχιση του 16S rRNA [11],[12]. Οι βασικές βελτιώσεις της εν λόγω τεχνικής είναι αρχικά η ανάλυση δειγμάτων χωρίς την ανάγκη κλωνοποίησης αλλά και η ταχύτητα αλληλούχισης των δειγμάτων. Μια από τις βασικές εφαρμογές της τεχνικής 16S rRNA Sequencing, είναι, ο προσδιορισμός της αλληλουχίας των προϊόντων ενίσχυσης (amplicons) του 16S rRNA σε σύνθετες μικροβιακές κοινότητες, όπου στη συνέχεια μας επιτρέπει τη σύγκριση των χλωρίδων μεταξύ διαφορετικών νοσολογικών καταστάσεων. Ένα κλασσικό παράδειγμα είναι η σύγκριση του μικροβιώματος του εντέρου μεταξύ αδύνατων και παχύσαρκων ατόμων [13] όπου έχει οδηγήσει στον εντοπισμό βακτηριακών ειδών που συσχετίζονται με την παχυσαρκία.

Η μεγαλύτερη ίσως προσφορά της τεχνικής 16S rRNA είναι πως μας φανέρωσε την πραγματική πολυπλοκότητα της ποικιλομορφίας του μικροβιώματος του ανθρώπου σε σχέση με αυτά που πιστεύαμε και γνωρίζαμε πριν.

1.8.5 Μαζική αλληλούχιση μεταγονιδιώματος (Shotgun metagenomic sequencing)

Γενικά, η μεταγονιδιωμική, ορίζεται ως η ανάλυση της σύστασης μιας μικροβιακής κοινότητας με πειραματικές και υπολογιστικές μεθόδους χωρίς την ανάγκη εργαστηριακής καλλιέργειας. Οι σύγχρονες μέθοδοι μεταγονιδιωμικής, συνήθως προσπαθούν να απαντήσουν στα ερωτήματα ποια μικρόβια υπάρχουν στο υπό μελέτη δείγμα και σε τι ποσότητα το καθένα από αυτά. Έχοντας αυτές τις πληροφορίες, μπορούν να εξαχθούν πολλά βιολογικά συμπεράσματα.

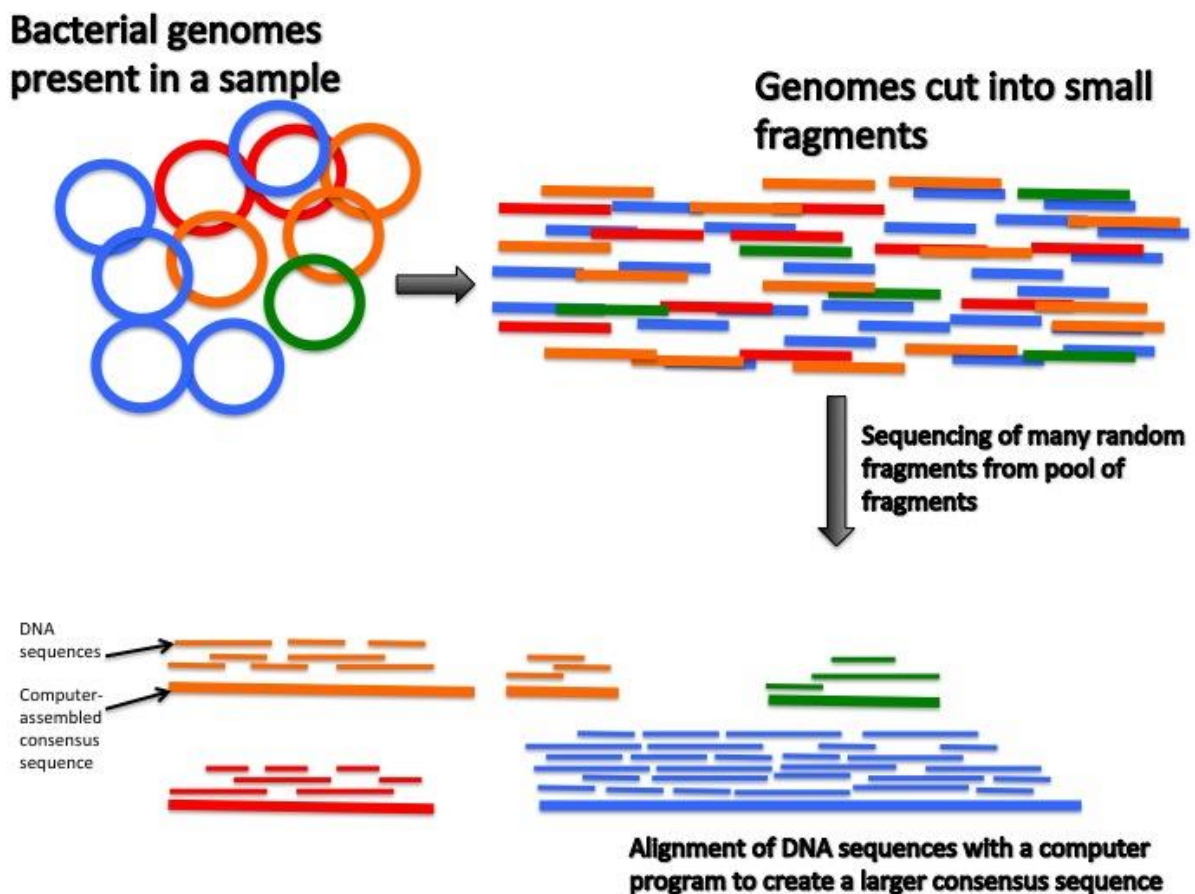
Από τότε που αναπτύχθηκε η τεχνολογία μαζικής αλληλούχισης του μικροβιώματος, έχει χρησιμοποιηθεί σε έρευνες μεγάλης εμβέλειας για την μελέτη του μικροβιώματος δειγμάτων από ωκεανούς [14], δειγμάτων από την εντερική χλωρίδα του ανθρώπου [15], [16], δειγμάτων χώματος [17], [18] και άλλα.

Η μαζική αλληλούχιση της μικροχλωρίδας ενός δείγματος, περιλαμβάνει τον προσδιορισμό της αλληλουχίας του γονιδιώματος του συνόλου των μικροβίων της. Για να επιτευχθεί αυτό, γίνεται μαζική, παράλληλη αλληλούχιση ολόκληρου του γονιδιώματος των μικροβίων που περιέχονται σε ένα δείγμα.

Τα βασικά βήματα της τεχνικής που περιγράφεται είναι αρχικά η εκχύλιση του γενετικού υλικού, στη συνέχεια η τυχαία τμηματοποίηση του γενετικού υλικού (DNA) του δείγματος, η αλληλούχιση των τμημάτων που προκύπτουν από το προηγούμενο βήμα και τέλος η σύνθεση των αλληλοκαλυπτόμενων περιοχών σε μια συνεχή ακολουθία (consensus sequence) με τη χρήση κατάλληλων υπολογιστικών μεθόδων.

Υπάρχουν πολλά πλεονεκτήματα της εν λόγω τεχνικής σε σχέση με όσες περιεγραφήκαν νωρίτερα. Αρχικά, τα δεδομένα που προκύπτουν έχουν πολύ μεγαλύτερο όγκο και

συνεπώς μπορούν να εξαχθούν περισσότερες πληροφορίες για το υπό μελέτη δείγμα. Επίσης, η τεχνική του Shotgun metagenomic sequencing μαζί με την τεχνική 16S rRNA-Seq είναι ταχύτερες σε σχέση με τις προηγούμενες τεχνικές και το κόστος τους συνεχώς μειώνεται. Παράλληλα, εκτός από την γενετική ποικιλότητα μιας μικροχλωρίδας, αποκαλύπτεται και το σύνολο των γονιδίων των μικροβίων που υπάρχουν και άρα μπορούμε να εξαγάγουμε συμπεράσματα και για τις λειτουργίες της μικροχλωρίδας μελετώντας και το μεταγράψωμα μεγάλων μικροβιακών κοινοτήτων (metatranscriptomics). Κάνοντας χρήση αυτών των πληροφοριών, μπορούμε να αποφανθούμε για το αν υπάρχει συσχέτιση του ανθρώπινου μικροβιώματος μεταξύ υγιών ή νοσογόνων καταστάσεων.



Εικόνα 4 - Τα βήματα της τεχνικής Shotgun Metagenomic Sequencing

1.9 State of the art

1.9.1 MetaPhlAn2

Το MetaPhlAn2 [21] [22] είναι ένα υπολογιστικό εργαλείο από το εργαστήριο “Segata Lab” του Πανεπιστημίου του Τρέντο το οποίο έχει ως κύριο στόχο το χαρακτηρισμό της σύνθεσης των μικροβιακών κοινοτήτων (βακτήρια, αρχαία, ιοί, ευκαρυωτικοί οργανισμοί) από δεδομένα μεταγονιδιωμικής που προέρχονται από πειράματα μαζικής αλληλούχησης μεταγονιδιώματος (shotgun metagenomic sequencing). Το εργαλείο παράγει αποτελέσματα κατά βάση σε επίπεδο ειδών (species) και σε σπάνιες περιπτώσεις και σε επίπεδο υποειδών (subspecies, strains).

Το MetaPhlAn2 εκτιμά τη σχετική αφθονία των μικροβιακών κοινοτήτων στοιχίζοντας (by aligning) τα reads του υπό ανάλυση δείγματος, έναντι μιας μειωμένης ομάδας ακολουθιών που χαρακτηρίζουν μονοσήμαντα ταξινομικούς βιολογικούς κλάδους (unique clade-specific sequence markers). Η ομάδα των ακολουθιών αυτών, έχει προεπιλεγεί υπολογιστικά από λειτουργικές περιοχές (coding sequences) των γονιδιωμάτων των μικροοργανισμών και η βασική ιδέα είναι πως εφόσον οι επιλεγμένες ακολουθίες χαρακτηρίζουν μονοσήμαντα έναν μικροβιακό οργανισμό ή μια ομάδα μικροβιακών οργανισμών, κάθε φορά που ένα read θα στοιχίζεται σε μια από τις ακολουθίες αυτές, αυτομάτως θα γνωρίζουμε από ποιο μικρόβιο ή ομάδα μικροβίων προέρχεται το read χωρίς την ανάγκη περαιτέρω ανάλυσης. Οι μικροβιακοί κλάδοι (microbial clades) που χρησιμοποιεί το MetaPhlAn2 μπορεί να είναι όσο γενική είναι η ταξινομική βαθμίδα του Φύλου (phylum) ή όσο ειδική είναι η ταξινομική βαθμίδα του Είδους. Σε εξαιρετικά σπάνιες περιπτώσεις, το MetaPhlAn2 μπορεί να εξάγει αποτελέσματα και για υποείδη χρησιμοποιώντας ακολουθίες που προέρχονται από τον μικροβιακό κλάδο των υποειδών.

Για το εν λόγω εργαλείο, οι βιολογικοί κλάδοι, είναι ομάδες γονιδιωμάτων μικροοργανισμών, οι οποίες μπορεί να είναι όσο ειδική είναι η ταξινομική βαθμίδα των ειδών ή όσο γενική είναι η ταξινομική βαθμίδα των φύλων. Για να εισαχθεί μια ακολουθία κάποιου γονιδιώματος σε ένα κλάδο, θα πρέπει να είναι συντηρημένη εντός των γονιδιωμάτων του κλάδου και να μην έχει σημαντικές τοπικές ομοιότητες με ακολουθίες γονιδιωμάτων εκτός του κλάδου.

Το MetaPhlAn2 στην τελευταία του έκδοση περιέχει περίπου ένα εκατομμύριο ακολουθίες (markers) και για την διαδικασία της στοίχισης των reads έναντι των ακολουθιών που περιέχονται στη βάση του, χρησιμοποιεί το εργαλείο Bowtie2.

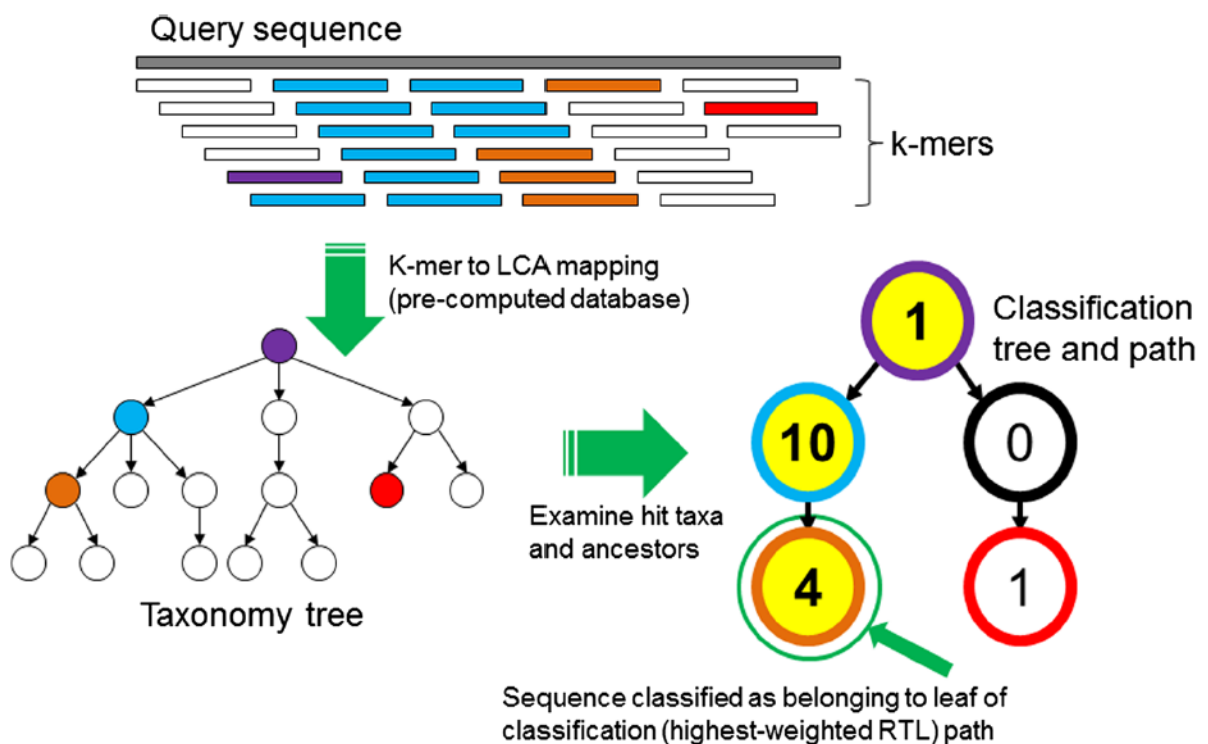
Όταν η διαδικασία της στοίχισης των reads του υπό εξέταση δείγματος ολοκληρωθεί, γίνεται κανονικοποίηση του συνολικού αριθμού των reads που στοιχήθηκαν (aligned) σε κάθε κλάδο με βάση το μέσο μήκος των ακολουθιών που περιέχονται στον κλάδο και έτσι προκύπτει η σχετική αφθονία καθενός από τους κλάδους.

Τα βασικά πλεονεκτήματα του MetaPhlAn2 έναντι άλλων προγραμμάτων που χρησιμοποιούνται σε δεδομένα μεταγονιδιωμικής είναι η υψηλή του ταχύτητα και οι πολλοί λίγοι υπολογιστικοί πόροι (RAM, CPU(s)) που χρειάζονται για να τρέξει το εργαλείο. Το βασικότερο ίσως μειονέκτημα της προσέγγισης αυτής είναι ότι συνήθως τα αποτελέσματα που προκύπτουν, καλύπτουν μόνο ένα μικρό ποσοστό (10% - 20%) του συνόλου των δεδομένων που αναλύονται. Αυτό συμβαίνει διότι, όπως αναφέρθηκε και προηγουμένως, η βάση των ακολουθιών που χρησιμοποιείται από το MetaPhlAn2 είναι αρκετά περιορισμένη σε σχέση με το εύρος των μικροβιακών γονιδιωμάτων που υπάρχουν συνολικά και προφανώς για να αποφανθεί για ένα read θα πρέπει το read αυτό, να “τυχαίνει” να πέσει σε κάποια από τις ακολουθίες της βάσης του.

1.9.2 Kraken

Το Kraken [23] είναι ένα υπολογιστικό εργαλείο όπου προέρχεται από το Κέντρο Υπολογιστικής Βιολογίας (Center for Computational Biology) του Πανεπιστημίου John Hopkins. Η βασική ιδιότητα του εργαλείου είναι η ανάθεση ταξινομικών ετικετών (taxonomic labels) σε μικρές ακολουθίες DNA οι οποίες συχνά προέρχονται από πειράματα μεταγονιδιωμικής.

Το Kraken δημιουργεί μια βάση που περιέχει εγγραφές που αποτελούνται από ένα k-mer και τον κατώτατο κοινό πρόγονο (Lower Common Ancestor - LCA) όλων των οργανισμών των οποίων το γονιδίωμα περιέχει αυτό το k-mer. Η χρήση αυτής της βάσης, επιτρέπει την γρήγορη αναζήτηση του πιο ειδικού κόμβου (specific node) στο ταξινομικό δέντρο που σχετίζεται με το υπό εξέταση k-mer.



Εικόνα 5 - Ο αλγόριθμος του Kraken

Για την κατηγοριοποίηση (classification) ενός read, το Kraken στοιχίζει κάθε k-mer του read στον κατώτατο κοινό πρόγονο (LCA) των γονιδιωμάτων που περιέχουν αυτό το k-mer. Στη συνέχεια, η Τάξη που σχετίζεται με τα k-mers του read, καθώς επίσης και οι πρόγονοι της τάξης, σχηματίζουν ένα υποδέντρο που έχει «κλαδευτεί» (pruned subtree) που προέρχεται από το γενικό ταξινομικό δέντρο και το οποίο χρησιμοποιείται για την κατηγοριοποίηση. Στο δέντρο κατηγοριοποίησης κάθε κόμβος έχει βάρος (weight) ίσο με τον αριθμό των k-mers που στοιχήθηκαν σε αυτή την τάξη. Τέλος, για κάθε μονοπάτι από τη ρίζα έως το φύλλο (root to leaf path - RTL) προστίθενται τα βάρη που βρίσκονται εντός των κόμβων και από το μεγαλύτερο άθροισμα προκύπτει η κατηγοριοποίηση του read η οποία γίνεται στο φύλλο του μονοπατιού.

Το γενικό ταξινομικό δέντρο των μικροοργανισμών που χρησιμοποιείται από το Kraken προέρχεται από το NCBI.

1.9.3 Clark

Το Clark [24] είναι κι αυτό ένα υπολογιστικό εργαλείο το οποίο κατηγοριοποιεί reads από πειράματα μεταγονιδιωμικής στις ταξινομικές βαθμίδες του γένους και των ειδών.

Σε πρώτη φάση, το εν λόγω εργαλείο παράγει όλα τα k-mers που προκύπτουν από όλα τα μικροβιακά γονιδιώματα που εμπεριέχονται στη βάση του εκάστοτε χρήστη. Στη συνέχεια, τα κοινά k-mers απομακρύνονται και η βάση του Clark παραμένει μόνο με τα μοναδικά k-mers (μοναδικές γονιδιωμικές περιοχές) που περιγράφουν μονοσήμαντα τους στόχους της βάσης του χρήστη. Κάθε read κατηγοριοποιείται στο γονιδίωμα στο οποίο στοιχίζεται ο μεγαλύτερος αριθμός από k-mers ενός read.

1.9.4 Bracken

Το Bracken [25] είναι ένας αλγόριθμος ο οποίος ουσιαστικά κάνει μετα-επεξεργασία (post-processing) στα αποτελέσματα του kraken. Το εν λόγω εργαλείο, εκτιμά τις αφθονίες των ειδών (species abundance estimation) σε δείγματα μεταγονιδιωμικής ακολουθώντας ένα πιθανοκρατικό μοντέλο μέσω του οποίου γίνεται ανακατανομή των reads στο ταξινομικό δέντρο (taxonomic tree) όπως αυτά έχουν προκύψει από το kraken.

Όπως είδαμε προηγουμένως, το Kraken, κατηγοριοποιεί κάθε read στον κατώτατο κοινό πρόγονο των γονιδιωμάτων που περιέχονται στη βάση του. Λόγω της υψηλής ομοιότητας πολλών ειδών (κυρίως στα βακτήρια), αυτό έχει ως αποτέλεσμα, πολλά από τα reads των υπό ανάλυση δεδομένων να κατηγοριοποιούνται στην ταξινομική βαθμίδα του γένους. Το Bracken, ακολουθώντας μια πιθανοκρατική προσέγγιση, χρησιμοποιεί τα αποτελέσματα του Kraken ώστε να κάνει ανακατανομή στα reads που είναι κατηγοριοποιημένα στη βαθμίδα του γένους, και να τα κατηγοριοποιήσει στη βαθμίδα τους ειδών. Επίσης, τα reads τα οποία είναι κατηγοριοποιημένα στην βαθμίδα των strains, κατηγοριοποιούνται κι αυτά στην βαθμίδα των species. Αυτό που πετυχαίνεται με αυτό τον τρόπο, είναι η εκτίμηση της αφθονίας των μικροβίων σε δεδομένα μεταγονιδιωμικής, σε επίπεδο ειδών. Η βελτιστοποίηση που προσφέρεται από το Bracken, καλύπτει ένα σημαντικό κενό καθώς όταν έχουμε τις αφθονίες των μικροβίων σε επίπεδο ειδών, μπορούν να εξαχθούν περισσότερα και ασφαλέστερα βιολογικά συμπεράσματα από το να έχουμε απλώς την κατηγοριοποίηση των reads στους κατώτατους κοινούς προγόνους. Το πιθανοκρατικό μοντέλο που χρησιμοποιείται από το εργαλείο, ακολουθεί Μπεϋζιανή (Bayesian) προσέγγιση.

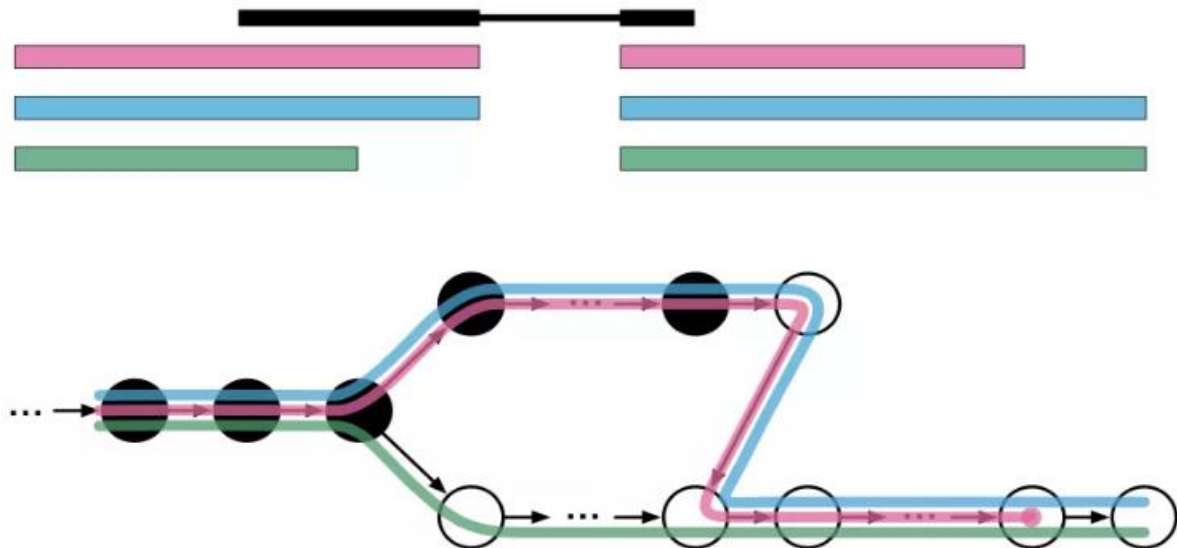
1.9.5 Kallisto

Το kallisto [26] είναι ένα υπολογιστικό εργαλείο το οποίο κάνει ποσοτικοποίηση (quantification) των transcripts σε πειράματα RNA-Seq.

Σε ένα pre-print [27] που εκδόθηκε, οι δημιουργοί του kallisto δείχνουν πως το εργαλείο μπορεί να ανταπεξέλθει και στο πρόβλημα της ποσοτικοποίησης των μικροβίων σε πειράματα μεταγονιδιωμικής.

Η γενική φιλοσοφία του αλγόριθμου είναι πως δεν είναι απαραίτητο να κάνουμε παραδοσιακή στοίχιση (traditional alignment) των reads ενός πειράματος στο γονιδίωμα που μας ενδιαφέρει για να ποσοτικοποιήσουμε τα transcripts με ακρίβεια, αρκεί να καθοριστεί, το πιο πιθανό transcript από το οποίο προέρχεται κάθε read.

Ο αλγόριθμος του Kallisto, αρχικά εξάγει όλα τα k-mers από το γονιδίωμα ή μεταγράψωμα και δημιουργεί ένα γράφο de Bruijn από τα k-mers αυτά, στη συνέχεια “τεμαχίζει” τα reads που προέρχονται από το πείραμα σε k-mers ίδιου μήκους με τα k-mers που περιέχονται στον γράφο de Bruijn και τέλος στοιχίζει τα k-mers των reads στον γράφο de Bruijn που έχει δημιουργηθεί. Η διαδικασία αυτή, ονομάστηκε από τους δημιουργούς του kallisto ως ψευδό-στοίχιση (pseudoalignment).



Εικόνα 6 - Η τεχνική του pseudo-alignment (kallisto)

Στην παραπάνω εικόνα φαίνεται σχηματικά η διαδικασία του pseudoalignment. Τα transcripts (που φαίνονται με χρώμα) τεμαχίζονται στα αντίστοιχα k-mers και στη συνέχεια χρησιμοποιούνται για να δημιουργήσουν ένα χρωματιστό Γράφο de Bruijn (colored de Bruijn graph) όπου κάθε transcript είναι κι ένα μονοπάτι στο γράφο και κάθε κόμβος είναι ένα k-mer του transcript. Ένα read (παρίσταται με μαύρο χρώμα) στοιχίζεται στο γράφο αντιστοιχίζοντας τα k-mers για να βρεθεί το πιο πιθανό transcript από το οποίο προέρχεται. Στην παραπάνω περίπτωση, το read προέρχεται είτε από το γαλάζιο είτε από το ροζ transcript.

Το kallisto χειρίζεται τα reads που στοιχίζονται σε πολλά transcripts (multi-alignments) χρησιμοποιώντας τον αλγόριθμο EM [28] και στο τέλος καταλήγει με κάθε read να έχει χρεωθεί σε ένα μοναδικό transcript.

1.10 Σκοπός

Σκοπός της παρούσας εργασίας, ήταν ο σχεδιασμός και η ανάπτυξη ενός υπολογιστικού εργαλείου το οποίο έχει τη δυνατότητα να εντοπίζει και να ποσοτικοποιεί τους οργανισμούς σε επίπεδο υποειδών (subspecies, strains) σε σύνθετα δείγματα μεταγονιδιωμικής και μικροβιώματος από πειράματα NGS.

Η ιδέα από την οποία ξεκίνησε η ανάπτυξη του εργαλείου, προήλθε από τα εξής: Αρχικά γνωρίζουμε από τη μοριακή βιολογία πως ο αριθμός των μικροοργανισμών που περιέχεται εντός και επί του ανθρωπίνου σώματος είναι πάρα πολύ μεγάλος. Υπάρχουν περιοχές στο ανθρώπινο σώμα με σχετικά μικρό αριθμό μικροβίων και όχι τόσο περίπλοκη ποικιλομορφία, ενώ άλλες, όπως το έντερο, περιέχουν τεράστιο αριθμό από μικρόβια με πολύ μεγάλο αριθμό διαφορετικών ειδών. Επίσης, γνωρίζουμε από σχετικές έρευνες, πως, υπάρχουν μικροοργανισμοί οι οποίοι είτε εισχωρούν εντός ανθρωπίνων επιθηλιακών κυττάρων που επενδύουν εσωτερικά διάφορες κοιλότητες είτε προσκολλώνται σε αυτές δημιουργώντας ένα επιπλέον στρώμα [29], [30], [31], [32].

Στα πειράματα αλληλούχισης επόμενης γενιάς τα οποία έχουν ως στόχο την αλληλούχιση του DNA ή RNA από δείγματα ιστού (tissue-specific samples) πολλές φορές γίνεται επιμόλυνση του δείγματος (sample contamination) κατά τη διαδικασία συλλογής, καλλιέργειας, ή προετοιμασίας αυτού [33], [34]. Τέλος, κατά την διαδικασία της στοίχισης των ακολουθιών του προέκυψαν από το πείραμα (reads), είναι σύνηθες, ένα ποσοστό των ακολουθιών να μην στοιχίζεται στο γονιδίωμα ή μεταγράφημα (transcriptome) του οργανισμού.

Λαμβάνοντας υπόψιν τα παραπάνω, το ερώτημα που γεννήθηκε ήταν το αν θα υπάρχει DNA/RNA μικροοργανισμών σε ανθρώπινα δείγματα ιστού ή σε δείγματα που προέρχονται από άλλους θηλαστικούς οργανισμούς και κατά πόσο θα μπορούσαμε να τα εντοπίσουμε.

Ιδανικά, χρησιμοποιώντας το εργαλείο που αναπτύχθηκε, θα μπορούσαν να βρεθούν μικρόβια σε δείγματα ιστού, και στη συνέχεια θα μπορούσαμε να εξετάσουμε δείγματα από υγιής περιοχές του ιστού και από περιοχές όπου έχει αναπτυχθεί κάποιος όγκος ή άλλη ασθένεια. Σε περίπτωση που βρίσκονταν μεγάλες διαφορές στις μικροβιακές κοινότητες των δειγμάτων ή σημαντική διαφορά σε συγκεκριμένα μικρόβια, θα μπορούσε να είναι η αρχή για να ξεκινήσουμε να ερευνούμε κατά πόσο το μικρόβιο ή μικρόβια αυτά, παίζουν κάποιο βιολογικό ρόλο στην ανάπτυξη της ασθένειας και άρα να συσχετίσουμε ασθένειες με μεμονωμένα μικρόβια ή και με ολόκληρες μικροβιακές κοινότητες.

Το εργαλείο που προέκυψε, έχει δύο βασικές ιδιότητες. Αρχικά μπορεί να χρησιμοποιηθεί ως αλγόριθμος κατηγοριοποίησης και ποσοτικοποίησης δεδομένων μεταγονιδιωμικής (metagenomics, metatranscriptomics) σε επίπεδο υποειδών και άρα και σε όλα τα ανώτερα επίπεδα όπως τα επίπεδα του είδους, του γένους, της οικογένειας και άλλα.

Η δεύτερη ιδιότητα είναι η χρήση του εργαλείου σε πειράματα γονιδιωμικής όπου ουσιαστικά έχουμε μικτά δείγματα που περιέχουν το DNA/RNA του ξενιστή και παράλληλα ψάχνουμε για DNA/RNA μικροβιακών οργανισμών. Για την συγκεκριμένη διαδικασία απαιτείται η εκτέλεση από μια σειρά από εργαλεία και αλγόριθμους όπου θα αναλυθούν στις μεθόδους.

Βασικοί στόχοι κατά την ανάπτυξη του εργαλείου ήταν η ανάλυση δειγμάτων με μεγάλη ακρίβεια σε επίπεδο υποειδών και με μεγάλη ταχύτητα καθώς τα δεδομένα μεταγονιδιωμικής εκτός από την δεδομένη πολυπλοκότητα τους, συνήθως είναι και αρκετά μεγάλου μεγέθους.

Ακόμα, ιδανικά το εργαλείο θα έπρεπε να είναι όσο το δυνατόν πιο απλό και εύκολο στη χρήση του, να μπορεί να εγκατασταθεί χωρίς προβλήματα, η ροή εργασιών του να είναι πλήρως αυτοματοποιημένη και τέλος να έχει την απαιτούμενη ελαστικότητα ώστε να μπορεί πολύ εύκολα να προσαρμοστεί στις ανάγκες του εκάστοτε ερευνητή ανάλογα με τους οργανισμούς που μελετάει και τα δείγματα που τον ενδιαφέρουν.

2. ΜΕΘΟΔΟΙ

2.1 To Pipeline

Για τον εντοπισμό και την ποσοτικοποίηση του μικροβιώματος σε δείγματα αλληλούχισης επόμενης γενιάς, αναπτύχθηκαν μια σειρά από προγράμματα. Αυτά, μαζί με ήδη υπάρχοντα εργαλεία, συνδυάστηκαν κατάλληλα και προέκυψε ένα pipeline το οποίο δέχεται ως είσοδο ένα αρχείο παραμέτρων που συμπληρώνεται από τον χρήστη καθώς και τα προς ανάλυση δείγματα και παράγει κάποια αρχεία εξόδου. Τα αρχεία αυτά, περιέχουν την αφθονία της μικροβιακής κοινότητας των δειγμάτων στις ταξινομικές βαθμίδες του γένους, των ειδών και των υποειδών.

Το pipeline είναι γραμμένο στη γλώσσα προγραμματισμού python, έκδοση 2.7 χρησιμοποιώντας την βιβλιοθήκη snakemake [35]. Με την συγκεκριμένη βιβλιοθήκη επιτρέπεται η δημιουργία κανόνων οι οποίοι περιγράφουν τον τρόπο με τον οποίο θα δημιουργηθούν τα αρχεία εξόδου από τα αρχεία εισόδου. Ουσιαστικά, για κάθε κανόνα περιγράφονται τα αρχεία εισόδου, οι διαδικασίες που θα εκτελεστούν, ο τρόπος και η σειρά με την οποία θα εκτελεστούν και τέλος τα αρχεία εξόδου που θα προκύψουν από την εκτέλεση του κανόνα. Αυτό μας δίνει τη δυνατότητα να χωρίσουμε το pipeline σε ξεχωριστές οντότητες (modules) όπου κάθε οντότητα περιγράφεται από ένα κανόνα. Έτσι, η υποστήριξη και ενημέρωση του pipeline γίνεται πολύ εύκολα καθώς για οποιαδήποτε μελλοντική αλλαγή, αρκεί να προστεθεί, ή να γίνει επεξεργασία κάποιου ή κάποιων modules, χωρίς να επηρεάζονται τα υπόλοιπα.

Ακολουθεί αναλυτική περιγραφή της ροής εργασιών που εκτελούνται από το pipeline.

2.2 Αρχείο παραμέτρων (configuration file)

Το αρχείο παραμέτρων είναι ένα αρχείο που συμπληρώνεται από τον χρήστη και περιέχει όλες τις παραμέτρους που είναι απαραίτητες για την εκτέλεση του pipeline. Μεταξύ αυτών είναι ο συνολικός αριθμός των επεξεργασιών, η μέγιστη μνήμη που μπορεί να δεσμεύσει το pipeline, η τοποθεσία των προς ανάλυση δειγμάτων κ.τ.λ.

Στη συνέχεια, κατά την εκτέλεση, υπολογίζεται ο αριθμός των επεξεργασιών που θα χρησιμοποιηθούν από κάθε διεργασία, ώστε οι κανόνες να εκτελούνται παράλληλα. Αυτό έχει ως αποτέλεσμα την αύξηση της ταχύτητας εκτέλεσης του pipeline, ειδικά όταν έχουμε περισσότερα από ένα δείγματα να αναλύσουμε.

2.3 Έλεγχος αρχείων εισόδου

Προτού ξεκινήσει η ανάλυση των δειγμάτων, γίνεται έλεγχος αυτών ώστε να διαπιστωθεί η ορθότητα τους για χρήση από το pipeline. Ο πρώτος έλεγχος αφορά το αν τα αρχεία εισόδου είναι FastQ ή FASTA, σε περίπτωση που έστω κι ένα από τα αρχεία δεν περάσει τον έλεγχο, ο χρήστης ενημερώνεται σχετικά και η εκτέλεση διακόπτεται.

Ο δεύτερος έλεγχος εφαρμόζεται μόνο στα paired-end δείγματα και ουσιαστικά ελέγχεται το αν κάθε read από το πρώτο αρχείο έχει και το αντίστοιχο ταίρι του (mate) στο δεύτερο αρχείο. Από όσα reads λείπει το ταίρι τους, διαγράφονται από το αρχείο. Ο συγκεκριμένος έλεγχος είναι απαραίτητος καθώς ο HISAT2 [36] που είναι το εργαλείο που

χρησιμοποιείται για την στοίχιση των reads στο γονιδίωμα ή μεταγράψωμα του ξενιστή, σε περίπτωση που από κάποιο read λείπει το ταίρι του, τερματίζει την εκτέλεση του ενημερώνοντας κατάλληλα τον χρήστη. Θα πρέπει να σημειωθεί πως η συγκεκριμένη προσέγγιση δεν είναι συχνή καθώς, συνήθως, ένα εργαλείο που κάνει στοίχιση ακολουθιών, όταν συμβαίνει το παραπάνω, ή προσπαθεί να στοίχισι το read ακόμα κι αν λείπει το ταίρι του, ή απλώς το προσπερνάει και συνεχίζει με τα υπόλοιπα.

Όταν τα προς ανάλυση δείγματα είναι ιστοειδικά (tissue-specific) και άρα προέρχονται από ιστό ανθρώπου ή άλλου θηλαστικού οργανισμού, το pipeline θα πρέπει να εκτελεστεί με την πρώτη κατάσταση (mode 1). Αυτό σημαίνει πως σε πρώτη φάση θα γίνει η στοίχιση των reads στο γονιδίωμα του ξενιστή και στην συνέχεια όσα reads δεν στοιχηθούν (unaligned), θα χρησιμοποιηθούν για τον εντοπισμό και την ποσοτικοποίηση του μικροβιώματος του ξενιστή. Σε περίπτωση που τα προς ανάλυση δείγματα προέρχονται από πειράματα μεταγονιδιωμικής, τότε χρησιμοποιείται η δεύτερη κατάσταση (mode 2) όπου περνάμε απευθείας στον εντοπισμό και ποσοτικοποίηση του μικροβιώματος.

2.4 Στοίχιση των δειγμάτων στο γονιδίωμα / μεταγράψωμα του ξενιστή

Η στοίχιση των reads των δειγμάτων στο γονιδίωμα ή μεταγράψωμα (genomes, transcriptome) του ξενιστή, γίνονται με το πρόγραμμα HISAT2. Το εν λόγω πρόγραμμα, έχει τη δυνατότητα να διενεργεί και spliced alignment.

Όταν τα προς ανάλυση δείγματα είναι DNA, η στοίχιση γίνεται στο γονιδίωμα του ξενιστή, ενώ όταν τα δείγματα είναι RNA στο μεταγράψωμα. Στην περίπτωση που τα δείγματα είναι RNA, εκτός της στοίχισης στο μεταγράψωμα, ο χρήστης μπορεί να επιλέξει να πραγματοποιηθεί και στοίχιση στο γονιδίωμα του ξενιστή.

Το αρχείο εξόδου του HISAT2 είναι ένα αρχείο SAM [37].

2.5 Απομάκρυνση των ακολουθιών του ξενιστή από το δείγμα

Μετά την εκτέλεση του HISAT2, ακολουθεί η απομάκρυνση των ακολουθιών που προέρχονται από τον ξενιστή.

Αυτό επιτυγχάνεται διαβάζοντας το SAM αρχείο που προέκυψε από τη διαδικασία της στοίχισης των ακολουθιών στον ξενιστή και βρίσκοντας τις ακολουθίες οι οποίες δεν στοιχήθηκαν σε καμία περιοχή του γονιδιώματος ή μεταγραφώματος (unaligned reads). Ο εντοπισμός των ακολουθιών αυτών πραγματοποιείται διαβάζοντας τον κωδικό του δεύτερου πεδίου από το SAM αρχείο και ελέγχοντας αν η τιμή κάθε σειράς υποδηλώνει ακολουθία που δεν έχει στοιχηθεί.

Κάθε read που εντοπίζεται, γράφεται σε ένα καινούργιο αρχείο SAM ακολουθώντας τη μορφή (format) των συγκεκριμένων αρχείων.

2.6 Απομάκρυνση του PhiX 174 από το δείγμα

Ο PhiX 174 βακτηριοφάγος είναι ένας ιός και ο πρώτος οργανισμός του οποίου έγινε η πλήρης αλληλούχιση του γονιδιώματος από τον Fred Sanger και την ομάδα του το 1977 [38].

Ο συγκεκριμένος οργανισμός, λόγω του μικρού μεγέθους του γονιδιώματος και του πολύ καλού χαρακτηρισμού του, χρησιμοποιείται από την εταιρεία Illumina ως ελεγκτής για πειράματα αλληλούχησης (control for sequencing runs – Phix Control V3). Αυτό έχει ως αποτέλεσμα, σε πολλά από τα δεδομένα που έχουν προκύψει από πειράματα με τις τεχνολογίες της Illumina, η συγκέντρωση του εν λόγω οργανισμού να είναι αρκετά μεγάλη.

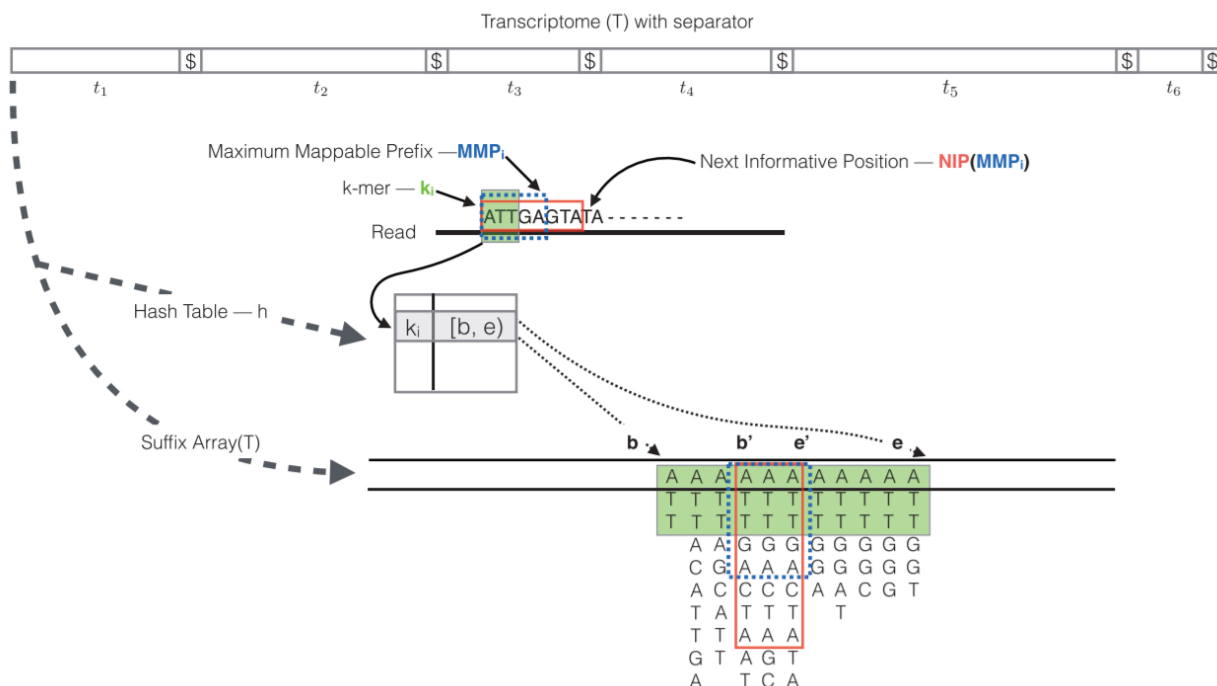
Έτσι, ένα επιπλέον βήμα πριν προχωρήσουμε στον εντοπισμό του μικροβιώματος είναι η απομάκρυνση των ακολουθιών που προέρχονται από τον οργανισμό PhiX 174. Ο τρόπος με τον οποίο επιτυγχάνεται αυτό, είναι ακριβώς όπως γίνεται και η απομάκρυνση των reads που προέρχονται από τον ξενιστή.

2.7 Εντοπισμός του μικροβιώματος

Για τον εντοπισμό του μικροβιώματος σε δείγματα NGS, χρησιμοποιείται ο αλγόριθμος του quasi-mapping όπως αυτός σχεδιάστηκε και υλοποιήθηκε από τον Rob Patro στο υπολογιστικό εργαλείο RapMap [39]. Ο συγκεκριμένος αλγόριθμος, στη συνέχεια υιοθετήθηκε από ακόμα δύο εργαλεία, το Salmon [40] και το Sailfish [41], και τα δύο του ίδιου δημιουργού.

Σε πρώτη φάση, ο αλγόριθμος του quasi-mapping, ενώνει όλα τα γονιδιώματα των μικροβίων που περιέχονται στο reference (αρχείο FASTA) του χρήστη, διαχωρίζοντας καθένα από αυτά με το σύμβολο του δολαρίου (\$-separated string). Στη συνέχεια, δημιουργείται ένας πίνακας καταλήξεων (Suffix Array - SA) από το αλφαριθμητικό που προέκυψε ενώνοντας όλα τα γονιδιώματα και παράλληλα δημιουργείται ένας πίνακας κατατεμαχισμού (hash-table) ο οποίος περιέχει όλα τα k-mers του αλφαριθμητικού και για κάθε k-mer το διάστημα στο SA στο οποίο εμφανίζεται αυτό το k-mer. Μιας και τα περιεχόμενα ενός SA είναι τοποθετημένα με λεξικογραφική σειρά, κάθε k-mer θα εμφανίζεται σε ένα συνεχές διάστημα εντός του SA.

Στη συνέχεια, για κάθε read που περιέχεται στα δεδομένα που επεξεργαζόμαστε, παράγονται όλα τα k-mers αυτού και εντοπίζεται το διάστημα στο SA στο οποίο περιέχεται το k-mer (αν υπάρχει), υπολογίζεται το μέγιστο πρόθεμα του read το οποίο στοιχίζεται στο SA (maximum mappable prefix - MMP) το οποίο πάντα ξεκινάει από το υπό εξέταση k-mer, και τέλος υπολογίζεται η επόμενη θέση που θα μας προσφέρει επιπλέον πληροφορία για τη στοίχιση του read (next informative position - NIP). Για τον εντοπισμό του NIP, υπολογίζεται το μέγιστο κοινό πρόθεμα (longest common prefix - LCP) μεταξύ των υπο-αλφαριθμητικών (substrings) και η NIP είναι το LCP + 1. Με αυτό τον τρόπο πετυχαίνουμε την παράκαμψη αρκετών k-mers του read (k-mer skipping) και άρα η στοίχιση των reads γίνεται πολύ πιο γρήγορα από μια τεχνική παραδοσιακής στοίχισης (alignment). Στην πραγματικότητα, για κάθε read που στοιχίζεται χρησιμοποιείται περίπου το 40 – 60% των k-mers που παράγονται από αυτό και συνεπώς από το σύνολο των δεδομένων μας, χρησιμοποιείται προσεγγιστικά περίπου το 50% αυτών.



Εικόνα 7 - Η τεχνική του quasi-mapping

2.8 Υπολογισμός αφθονίας στις ταξινομικές βαθμίδες του γένους, είδους και υποείδους

Όταν γίνεται ο εντοπισμός και ποσοτικοποίηση των μικροβίων, οι αφθονίες που προκύπτουν αναφέρονται σε κάθε ένα γονιδίωμα ή πλασμίδιο ξεχωριστά. Όμως, το τελικό αποτέλεσμα που θέλουμε από ένα εργαλείο που κάνει ποσοτικοποίηση μικροβιώματος, είναι η αφθονία στο σύνολο των ταξινομικών βαθμίδων που μας ενδιαφέρουν.

Για να επιτευχθεί αυτό, γίνεται χρήση της ταξινομίας που προσφέρεται από το NCBI και που ουσιαστικά περιέχει για τα περισσότερα από τα γονιδιώματα, σε ποιο είδος, γένος, οικογένεια κ.τ.λ. ανήκουν.

Η ταξινομία του NCBI είναι οργανωμένη σε ένα δένδρο όπου για κάθε ένα από τα γονιδιώματα, ξεκινώντας από τις ρίζες του δένδρου, ακολουθείται μια διαδρομή και εξαγάγουμε τις ταξινομικές βαθμίδες του γονιδιώματος που ελέγχεται. Υπάρχουν μικροοργανισμοί για τους οποίους κάποιες βαθμίδες δεν είναι γνωστές, σε αυτή την περίπτωση ενημερώνεται σχετικά ο χρήστης στα αρχεία εξόδου.

Οι ταξινομικές βαθμίδες στις οποίες ανήκουν όλοι οι μικροβιακοί οργανισμοί που περιέχονται στο NCBI, είναι υπολογισμένες από πριν και ο υπολογισμός αυτός δεν χρειάζεται να επαναλαμβάνεται. Έτσι, έχοντας την πληροφορία αυτή, μπορούμε να αθροίζουμε τον αριθμό των reads κάθε γονιδιώματος στο γένος, είδος και υποείδος που ανήκει και τελικώς να προκύψουν η συνολική αφθονία των βαθμίδων που μας ενδιαφέρουν.

2.9 Φιλτράρισμα των αποτελεσμάτων

Όταν έχουμε τις αφθονίες στα επίπεδα του γένους, των ειδών και υποειδών, πολλές από τα αφθονίες αυτές μπορεί να είναι μηδενικές ή πολύ μικρές και να μην έχουν πρακτική αξία για τον ερευνητή. Για το λόγο αυτό, παρέχεται η δυνατότητα να δίνεται ένα όριο (cut-off) και οποιοδήποτε γένος, είδος, υποείδος έχει αφθονία μικρότερη του ορίου, να απομακρύνεται από τα τελικά αποτελέσματα.

Με αυτό τον τρόπο, μπορούν να δημιουργηθούν αρχεία τα οποία περιέχουν μόνο τα μικρόβια με πολύ μεγάλη αφθονία ή με μεγάλη και μεσαία. Φυσικά, υπάρχει και η δυνατότητα να περιέχονται όλα τα μικρόβια στα αποτελέσματα ανεξαρτήτως της αφθονίας τους στο δείγμα, απλά επιλέγονται το όριο ίσο με μηδέν.

2.10 Υπολογισμός βασικών στατιστικών

Σε αυτό το βήμα, γίνεται ο υπολογισμός μερικών βασικών στατιστικών που αφορούν τα αποτελέσματα της ανάλυσης.

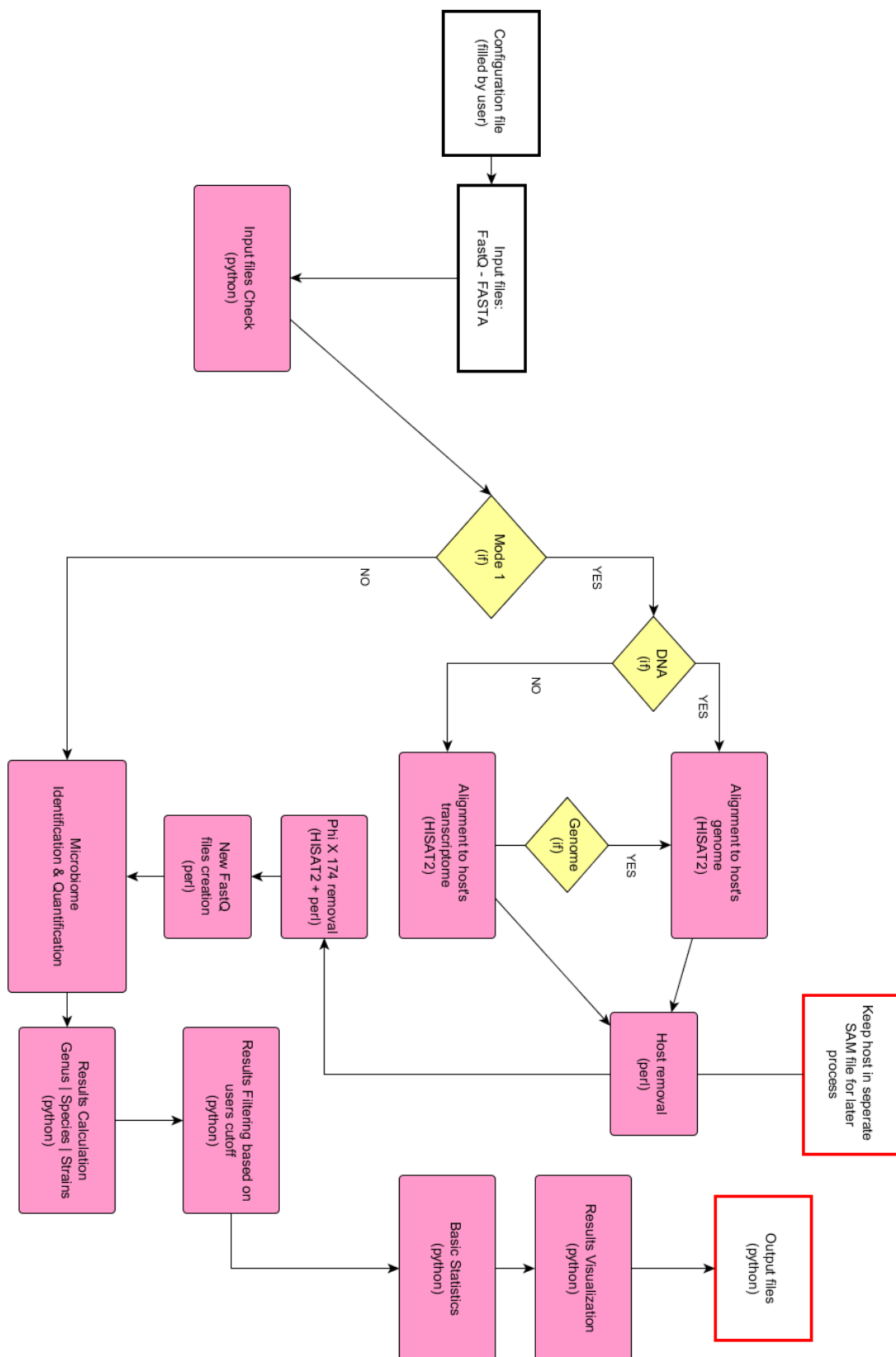
Εξάγεται το ποσοστό (επί της εκατό) των βακτηρίων, ιών και αρχαίων στο δείγμα, παρουσιάζονται τα δέκα γένη και είδη με τη μεγαλύτερη αφθονία, τα είκοσι υποείδη με τη μεγαλύτερη αφθονία και τέλος υπολογίζεται το συνολικό ποσοστό του μικροβιώματος σε σχέση με το συνολικό δείγμα.

2.11 Οπτικοποίηση αποτελεσμάτων

Μετά τον υπολογισμό των αποτελεσμάτων και βασικών στατιστικών, χρησιμοποιείται η βιβλιοθήκη matplotlib [42] για την οπτικοποίηση αυτών. Ουσιαστικά, δημιουργούνται γραφικές παραστάσεις, οι οποίες παρουσιάζουν τα παραπάνω αποτελέσματα.

2.12 Αρχεία εξόδου

Τα αρχεία εξόδου που προκύπτουν από την εκτέλεση του pipeline, είναι τρία ξεχωριστά text αρχεία με τις μικροβιακές αφθονίες στα επίπεδα του γένους, των ειδών και των υποειδών, ένα επιπλέον αρχείο όπου περιέχονται τα μικρόβια για τα οποία λείπει η πληροφορία της ταξινόμιάς τους, και τέλος, σε περίπτωση που αναλύθηκαν ιστοειδικά δείγματα, εξάγεται και ένα BAM αρχείο το οποίο περιέχει τα αποτελέσματα της στοίχισης στο γονιδίωμα του ξενιστή για περαιτέρω ανάλυση σε περίπτωση που ο ερευνητής το επιθυμεί. Η εξαγωγή του BAM αρχείου είναι προληπτική και γίνεται μόνον εφόσον ο χρήστης το επιλέξει.



Εικόνα 8 - Το Pipeline σχηματικά

3. ΣΥΝΘΕΤΙΚΑ ΔΕΔΟΜΕΝΑ

3.1 Δημιουργία συνθετικών δεδομένων

Ο πιο συνηθισμένος τρόπος για την αξιολόγηση της ακρίβειας ενός υπολογιστικού εργαλείου, είναι χρησιμοποιώντας συνθετικά δεδομένα για τα οποία γνωρίζουμε από πριν τις πραγματικές αφθονίες των δεδομένων μας ή ακόμα πιο λεπτομερώς, γνωρίζουμε για κάθε read, την περιοχή του γονιδιώματος από την οποία προέρχεται (ground truth).

Εκτός της προγενέστερης γνώσης για τα δεδομένα, η προσέγγιση αυτή έχει και μερικά ακόμα πλεονεκτήματα. Αρχικά, μπορούμε να δημιουργήσουμε σύνολα δεδομένων οποιουδήποτε μεγέθους επιθυμούμε, επίσης μπορούμε να κάνουμε τα δεδομένα μας όσο περίπλοκα θέλουμε. Τέλος, τα περισσότερα προγράμματα παρέχουν τη δυνατότητα επιλογής του αριθμού των μεταλλάξεων και του λάθους αλληλούχισης στα reads [43] που παράγουν και άρα έχουμε τη δυνατότητα να αξιολογήσουμε το εργαλείο που μας ενδιαφέρει σε διαφορετικές συνθήκες.

Έχοντας αναπτύξει το εργαλείο κατηγοριοποίησης και ποσοτικοποίησης των μικροβιακών κοινοτήτων σε δείγματα μεταγονιδιωμικής, η επόμενη ενέργεια ήταν η αξιολόγηση των αποτελεσμάτων του εργαλείου με τη χρήση συνθετικών δεδομένων.

Για τη δημιουργία των συνθετικών δεδομένων χρησιμοποιήθηκε το Art [44] το οποίο είναι ένα σύνολο από εργαλεία μέσω των οποίων παράγονται συνθετικά δεδομένα αλληλούχισης επόμενης γενιάς. Το Art, χρησιμοποιεί εμπειρικό μοντέλο για τον υπολογισμό του σφάλματος αλληλούχισης που θα έχει κάθε read (empirical error model) και το λάθος είναι ειδικό για κάθε μια από τις τεχνολογίες που υποστηρίζει. Από το πρόγραμμα υποστηρίζονται και διάφορες μεγάλες τεχνολογίες αλληλούχισης Illumina Solexa, Roche 454 και Applied Biosystems SOLiD. Τέλος, μια από τις σημαντικές ιδιότητες του εργαλείου, είναι πως επιτρέπει την δημιουργία μοντέλων σφάλματος από το χρήστη.

Για την αξιολόγηση του εργαλείου που αναπτύχθηκε στα πλαίσια της διπλωματικής, αλλά και για τη σύγκριση της ακρίβειας του εργαλείου με κάποια από τα δημοφιλή εργαλεία της ίδιας κατηγορίας, δημιουργήθηκε πληθώρα από συνθετικά σύνολα δεδομένων, πέντε από τα οποία θα παρουσιαστούν παρακάτω. Κάθε ένα από αυτά, είναι διαφορετικού μεγέθους, πολυπλοκότητας και άλλων τεχνικών προδιαγραφών.

Ο αριθμός των reads που παράχθηκαν από κάθε γονιδίωμα, καθώς και οι περιοχές των γονιδιωμάτων από τις οποίες προήλθε το κάθε read επιλέχθηκαν ψευδοτυχαία.

Πίνακας 1 - Τα συνθετικά δεδομένα που δημιουργήθηκαν

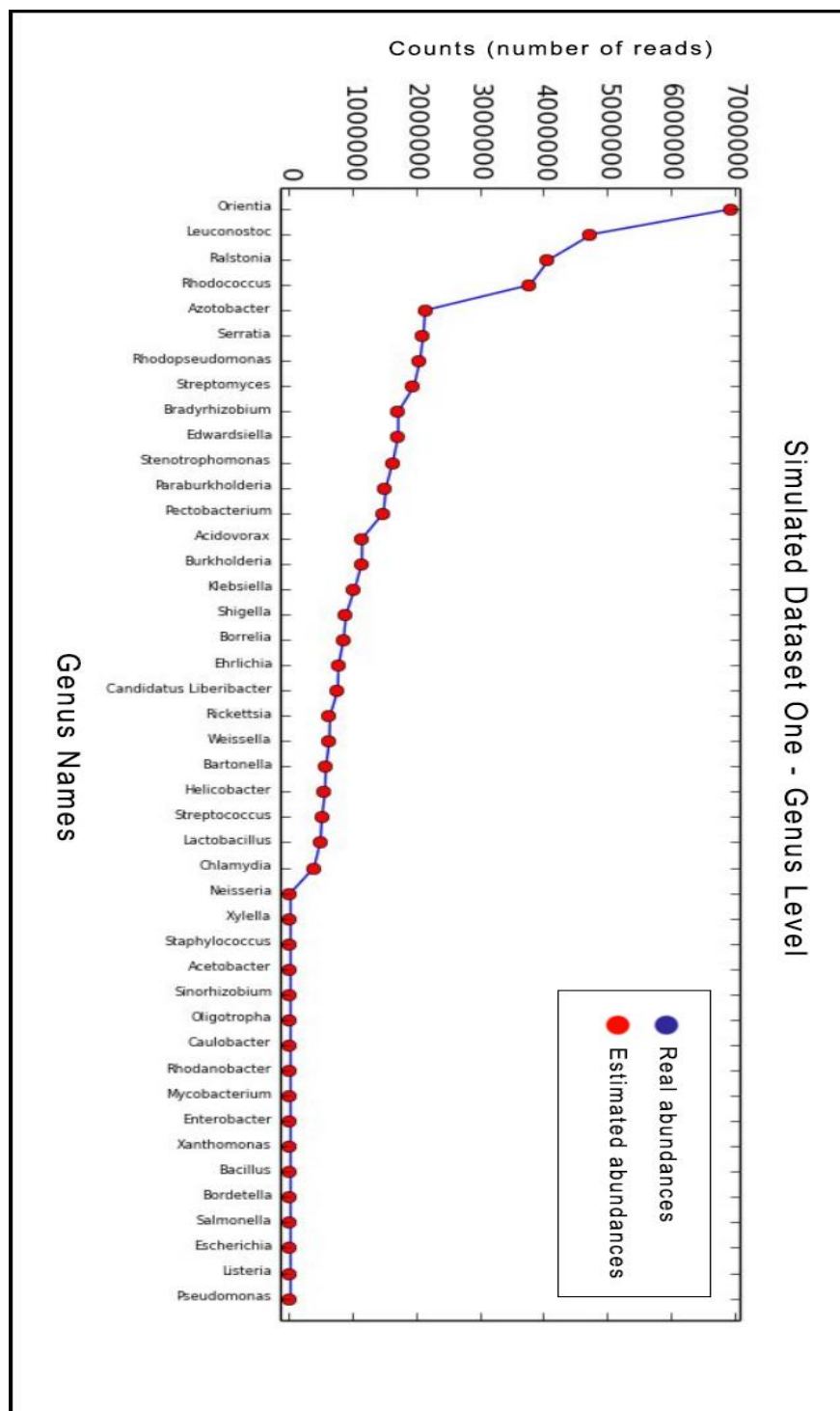
Όνομα dataset	Μέγεθος	Τεχνολογία	Μέγεθος των reads	Στρατηγική	Πολυπλοκότητα
sim_dataset1	45.7M reads	HiSeq 2500	100 bp	paired-end	35 genomes
sim_dataset2	97.6M reads	HiSeq 2000	76 bp	paired-end	65 genomes
sim_dataset3	69.3M reads	HiSeq 2000	100 bp	paired-end	61 genomes
sim_dataset4	153.4M reads	HiSeq 2500	100 bp	paired-end	130 genomes
sim_dataset5	415.3M reads	HiSeq 2500	100 bp	paired-end	625 genomes

3.2 Αποτελέσματα του metaHost

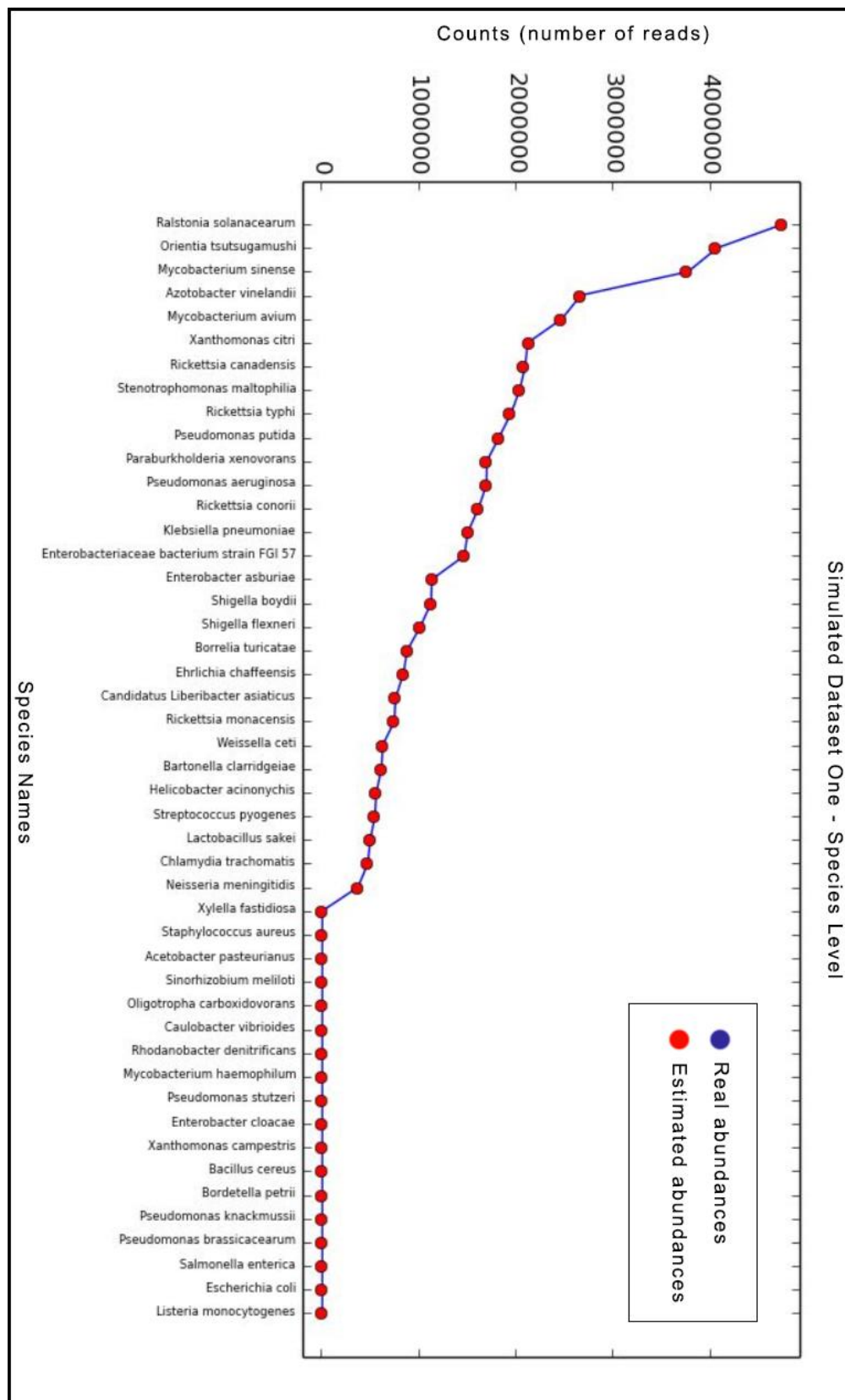
Τα αποτελέσματα που παρουσιάζονται παρακάτω μέσω γραφικών παραστάσεων, αφορούν τις ταξινομικές βαθμίδες του γένους (genus), του είδους (species) και του υποείδους (strains, subspecies). Σε κάθε γραφική παράσταση, απεικονίζονται η πραγματική αφθονία των μικροβίων με μπλε γραμμή, και οι αφθονίες που υπολόγισε το meta-host για κάθε μικρόβιο με κόκκινους κύκλους.

Στον άξονα των x παρέχονται τα ονόματα του γένους, είδους και υποείδους που περιέχονται στο κάθε σύνολο δεδομένων αντίστοιχα, ενώ ο άξονας των y αναφέρεται στον αριθμό των reads που προέρχονται από κάθε γένος, είδος και υποείδος αντίστοιχα.

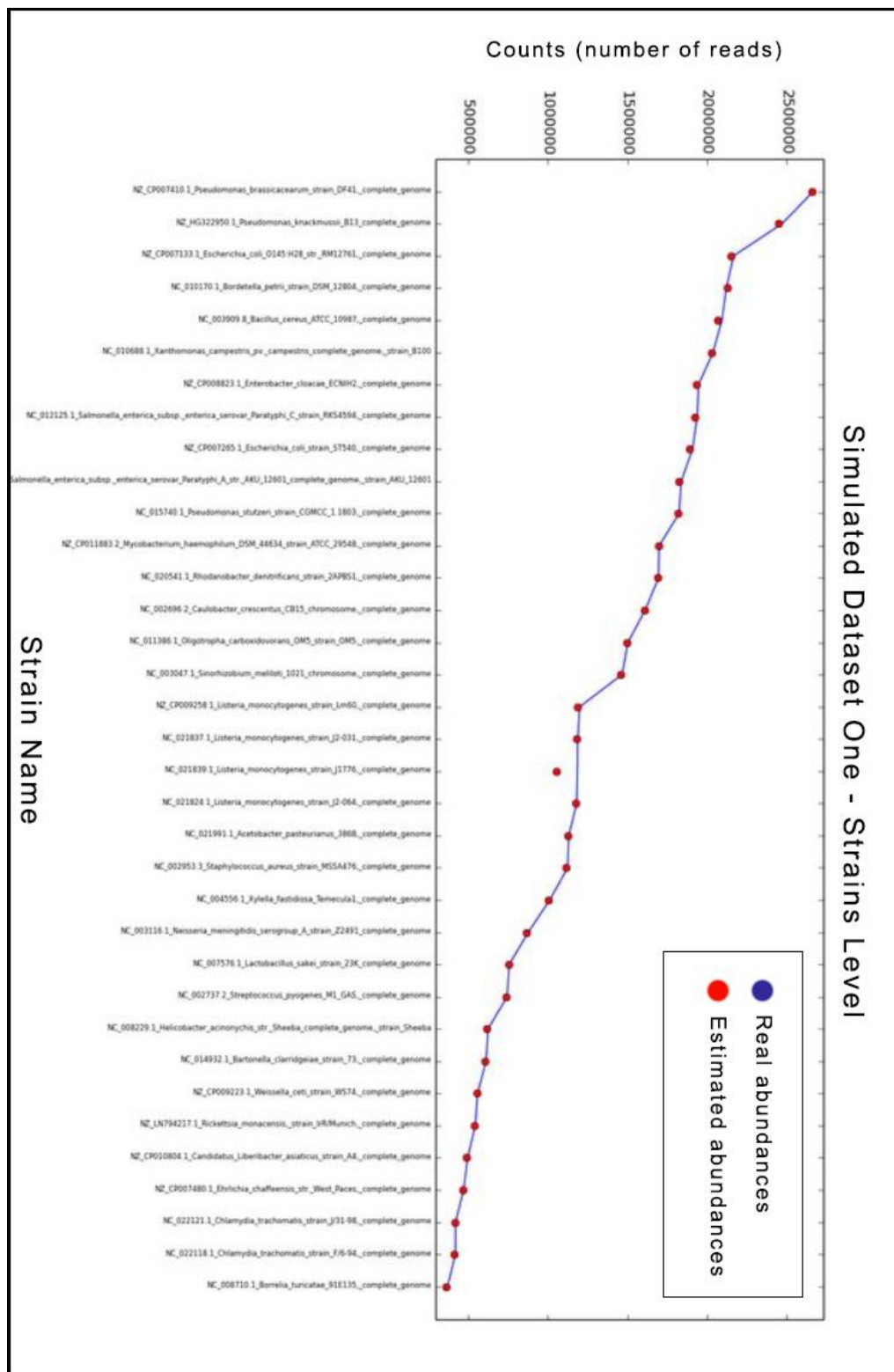
Αρχικά παρουσιάζονται τα αποτελέσματα του metaHost σε σχέση με τις πραγματικές αφθονίες των μικροβίων και στην συνέχεια, για τα ίδια δεδομένα, παρουσιάζονται τα αποτελέσματα του metaHost μαζί με τα αποτελέσματα κάποιων state-of-the-art προγραμμάτων με σκοπό την σύγκριση αυτών. Ακολουθούν οι γραφικές παραστάσεις των αποτελεσμάτων.



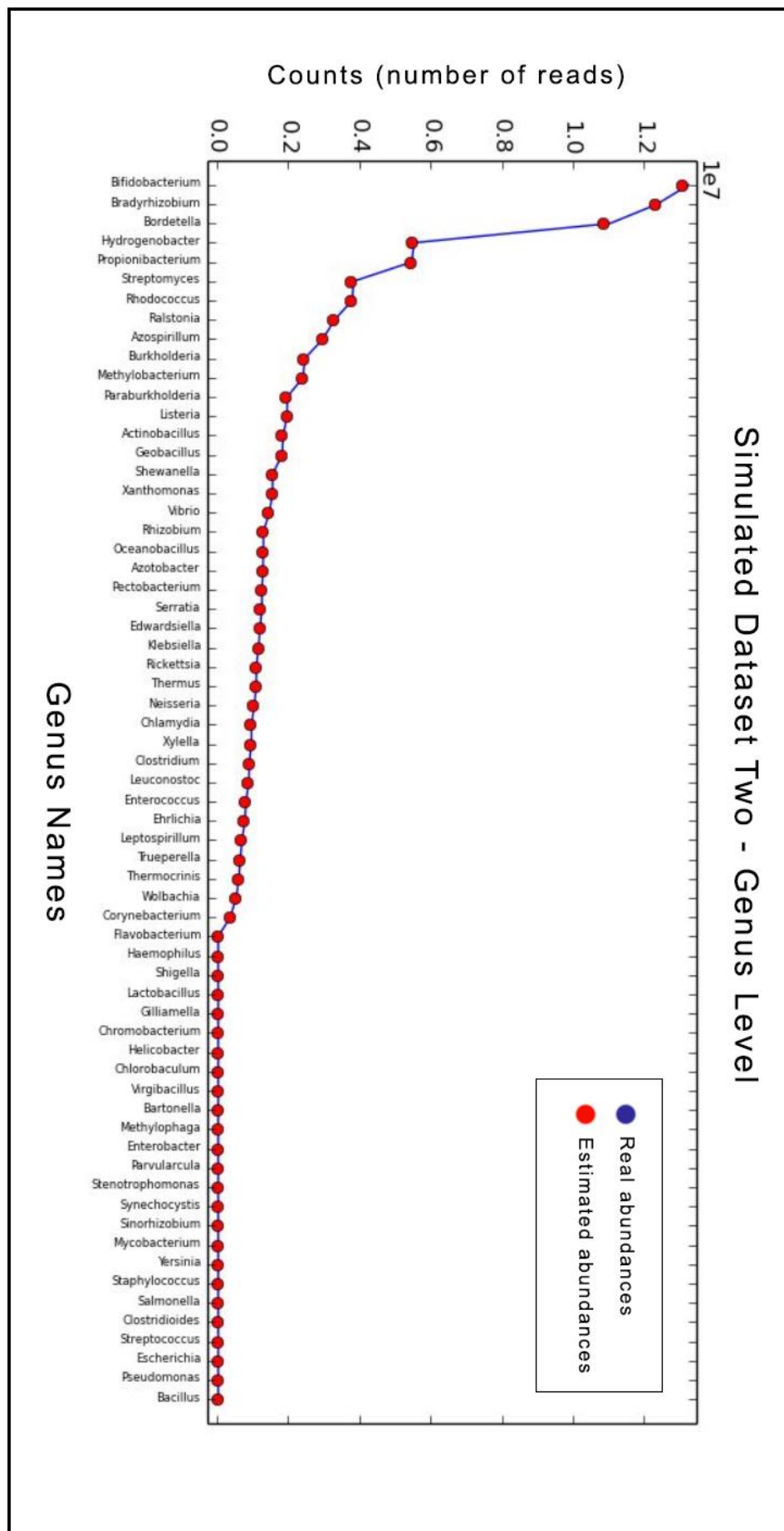
Σχήμα 1 - metaHost Dataset 1 - Genus level



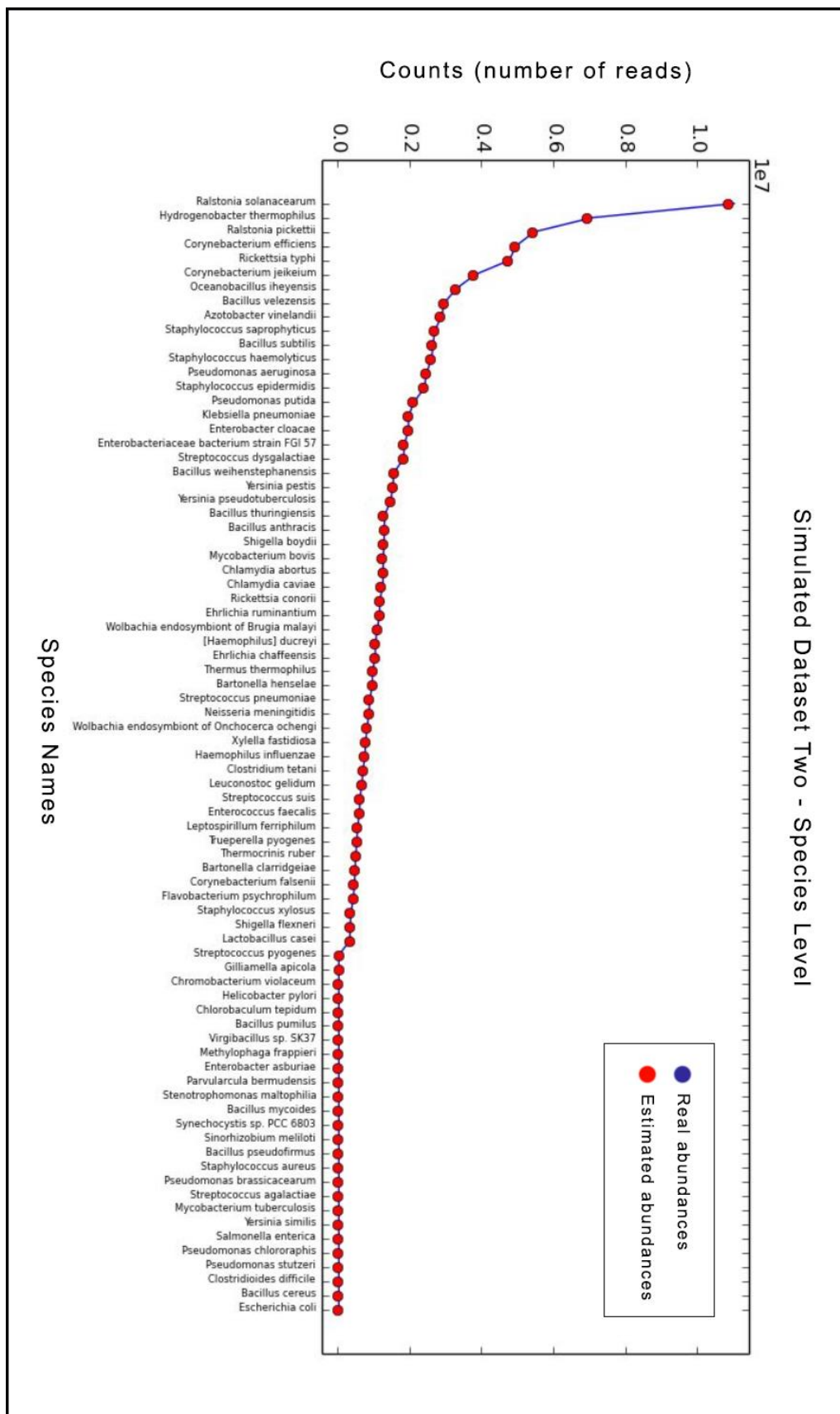
Σχήμα 2 - metaHost Dataset 1 - Species level



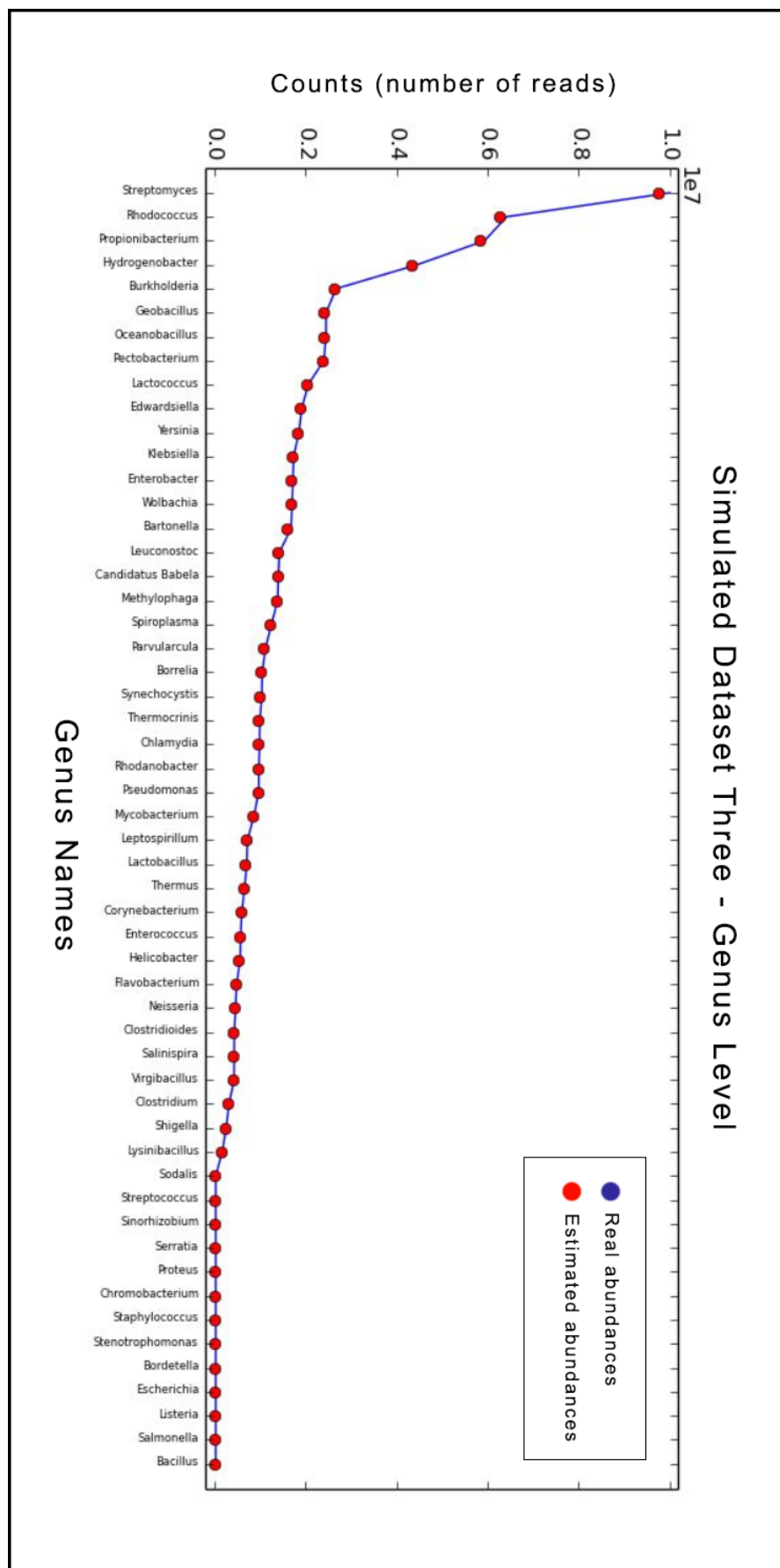
Σχήμα 3 - metaHost Dataset 1 - Strain level



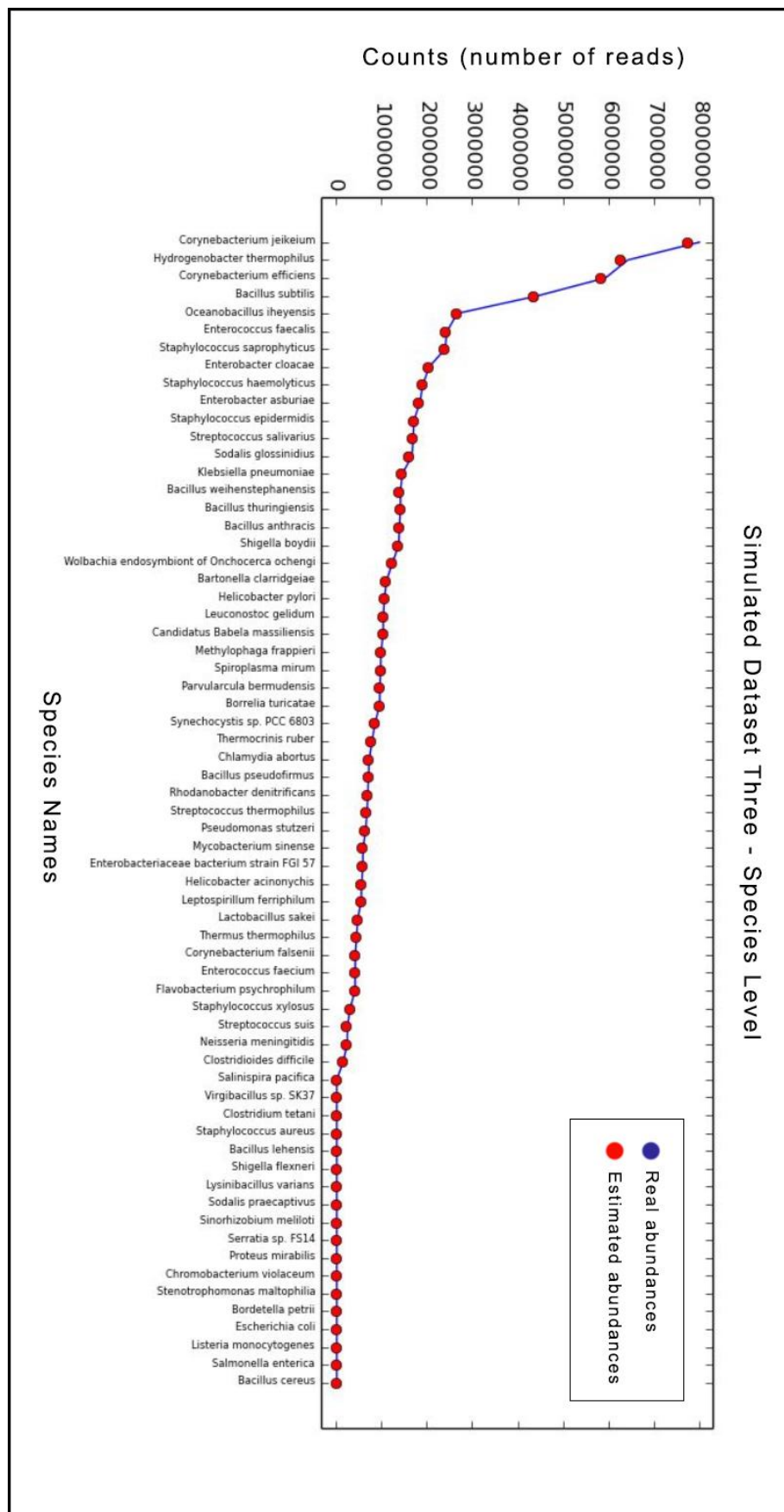
Σχήμα 4 - metaHost Dataset 2 - Genus level



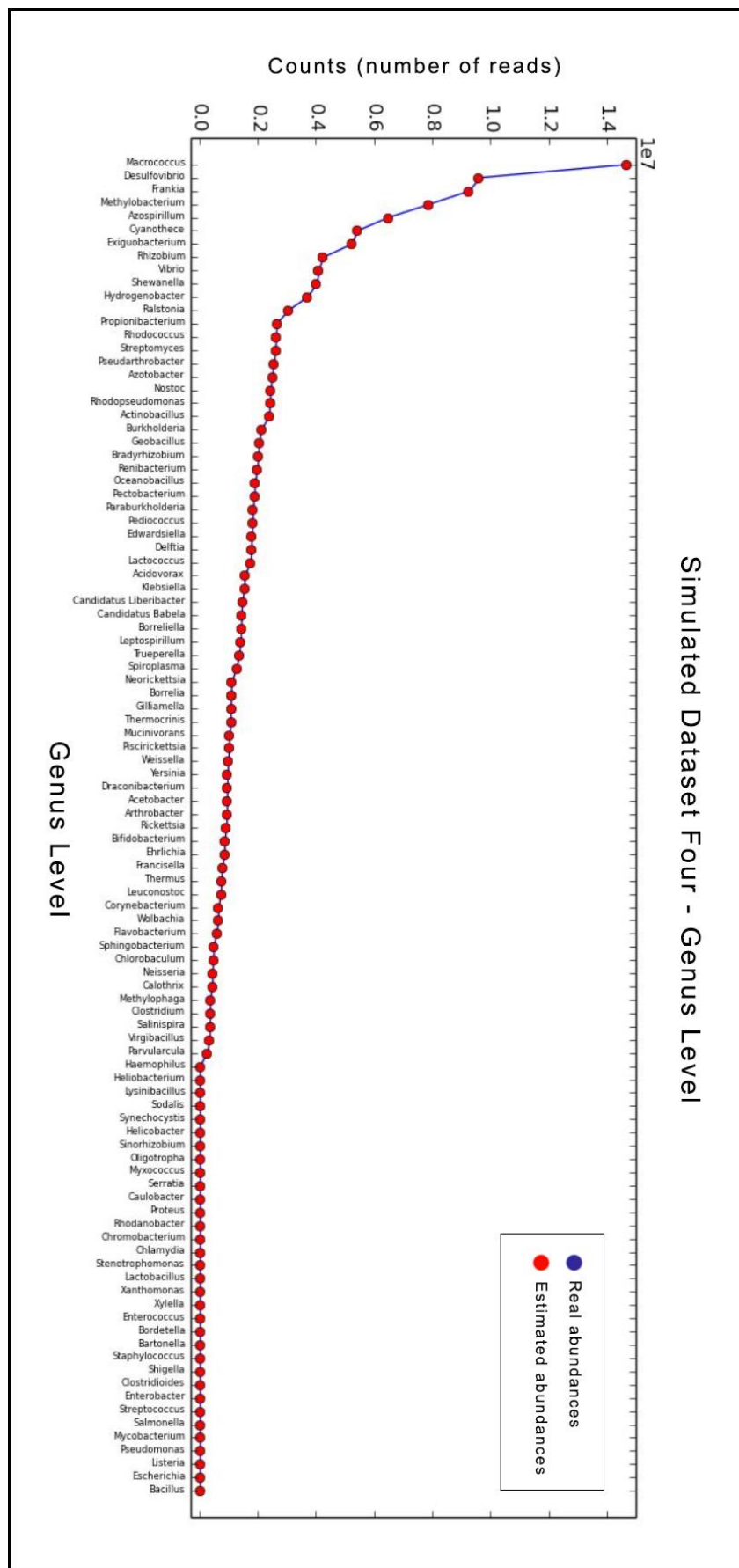
Σχήμα 5 - metaHost - Dataset 2 - Species level



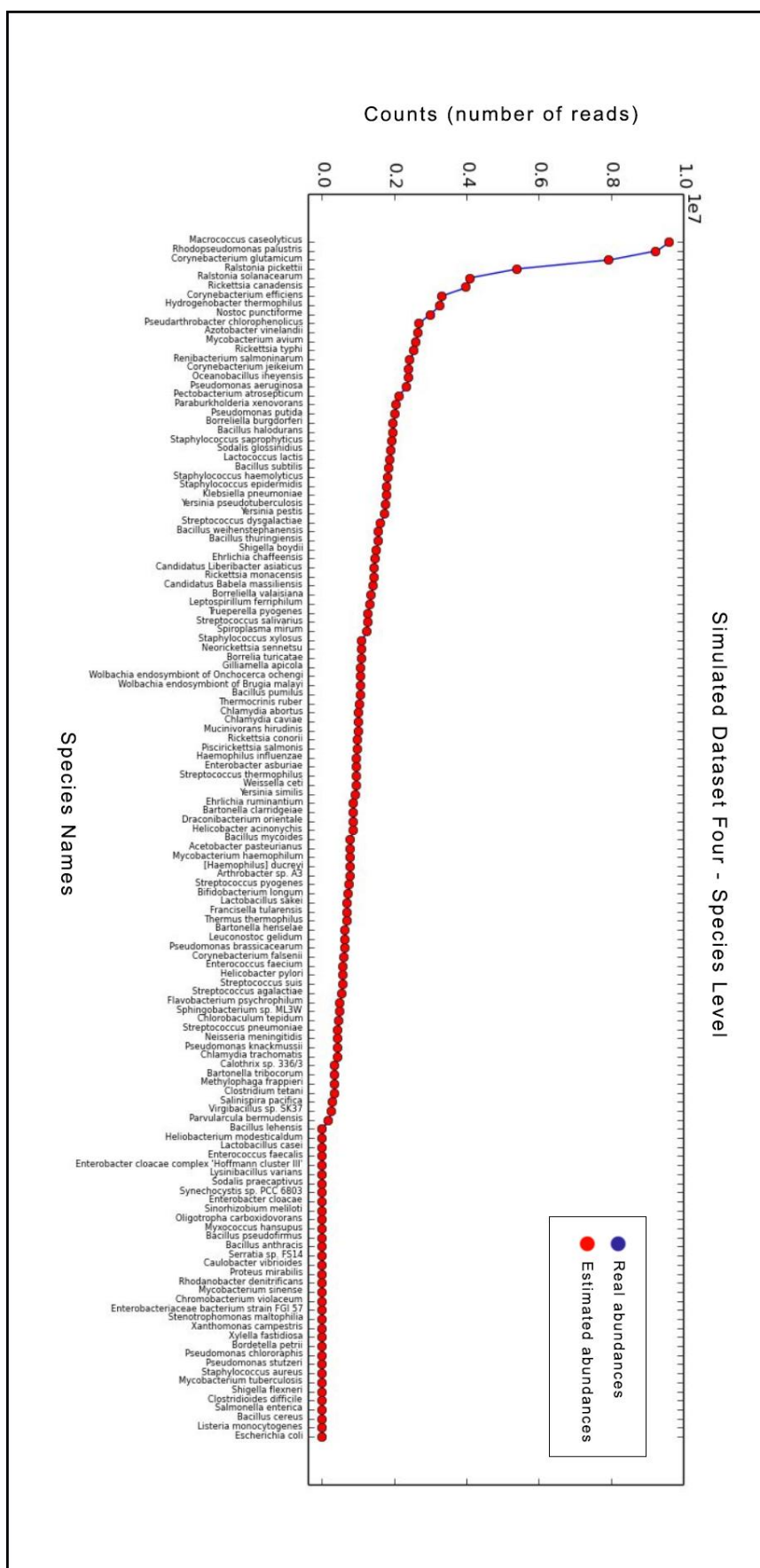
Σχήμα 6 - metaHost Dataset 3 - Genus level



Σχήμα 7 - metaHost Dataset 3 - Species level

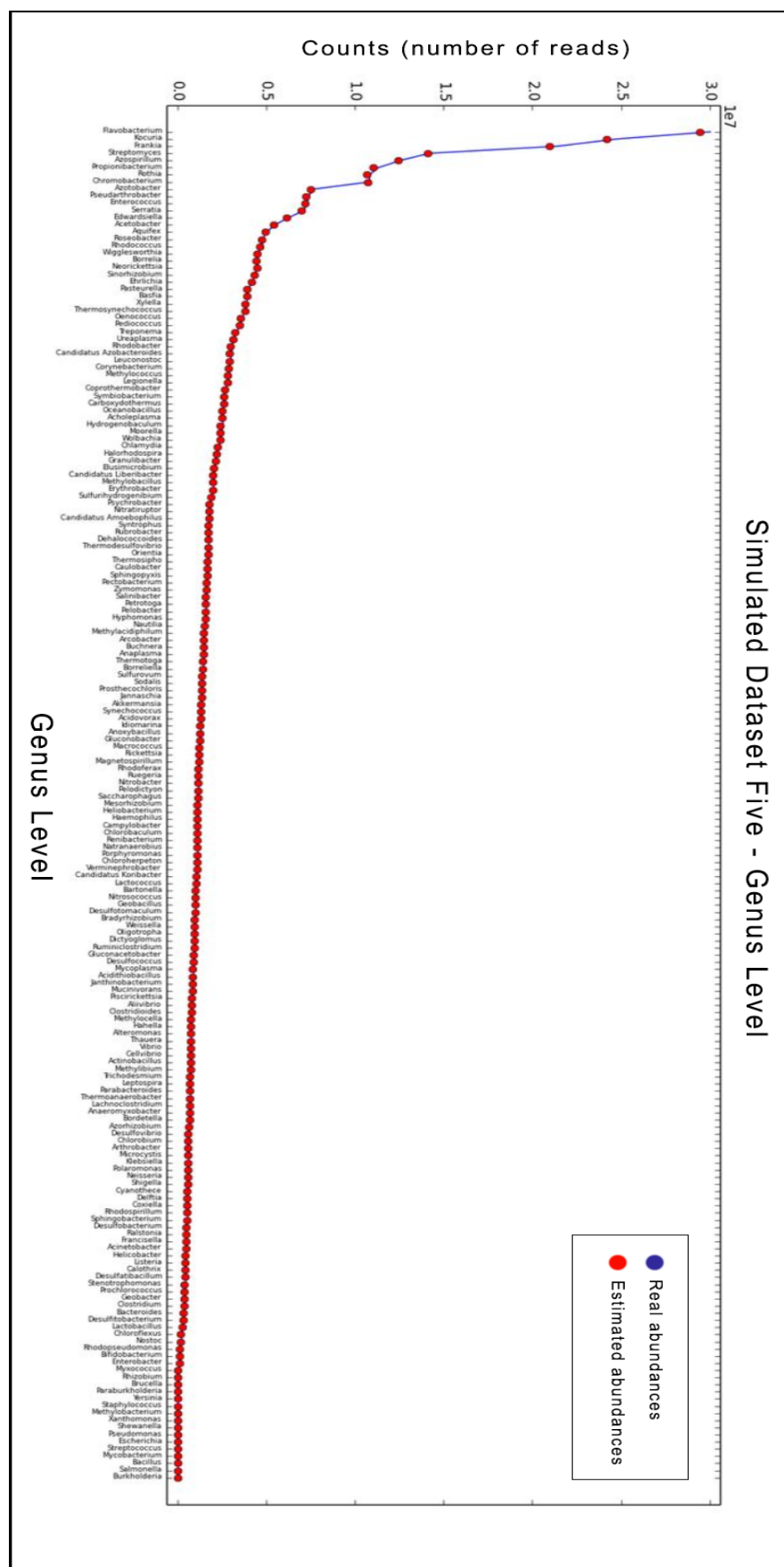


Σχήμα 8 - metaHost Dataset 4 - Genus level

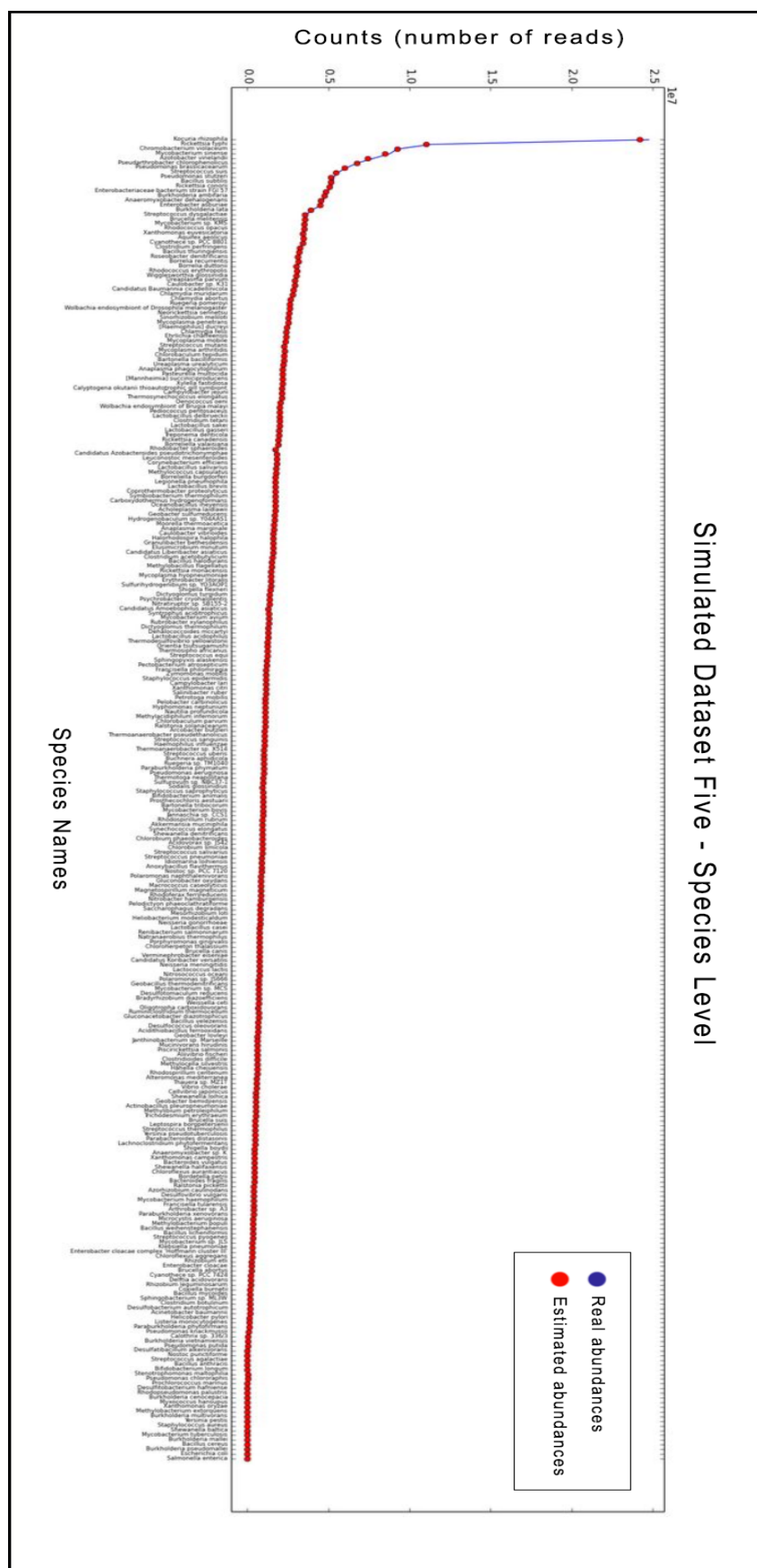


Σχήμα 9 - metaHost Dataset 4 - Species level

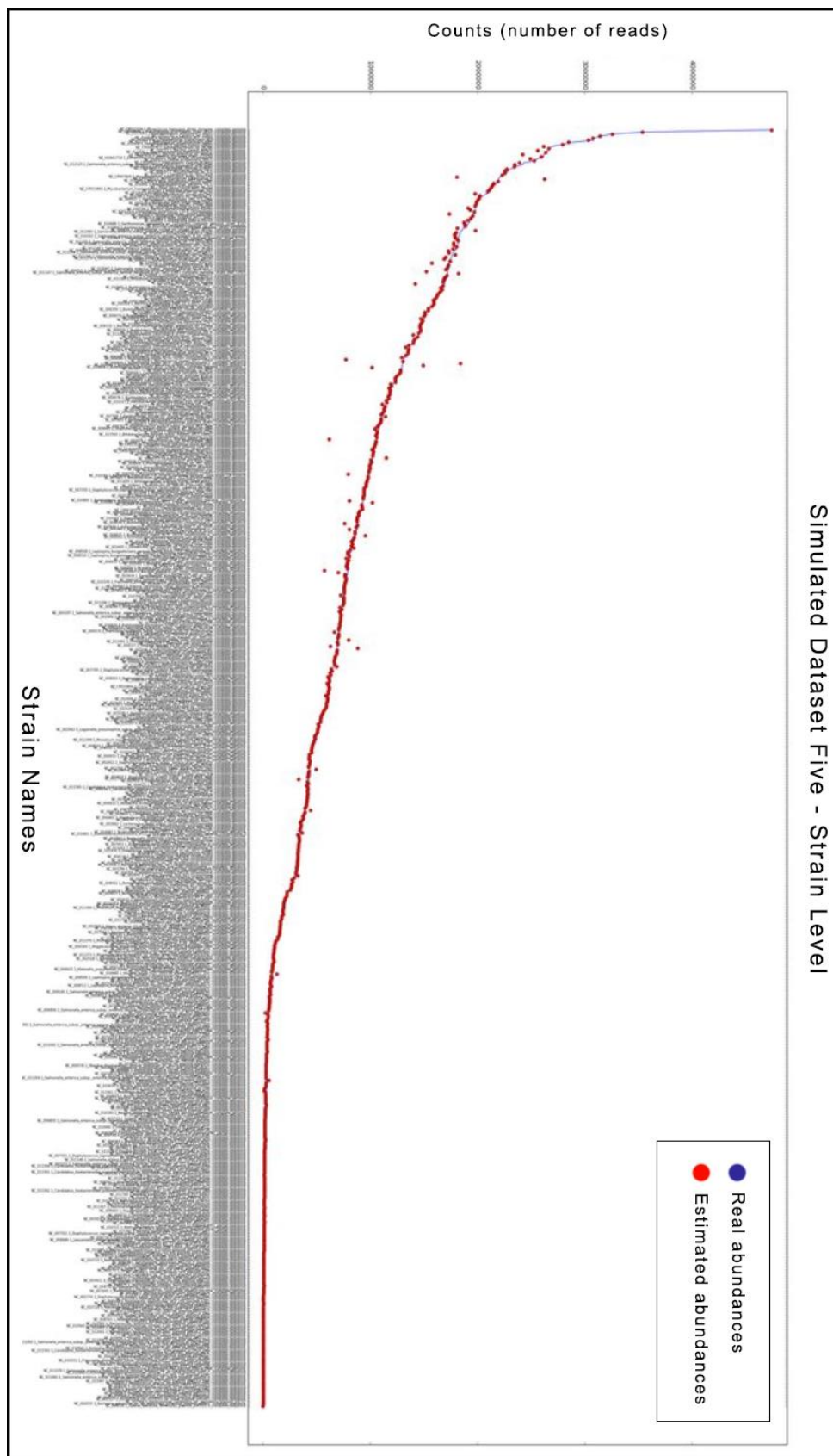
Σχήμα 10 - metaHost Dataset 4 - Strain level



Σχήμα 11 - metaHost Dataset 5 - Genus level



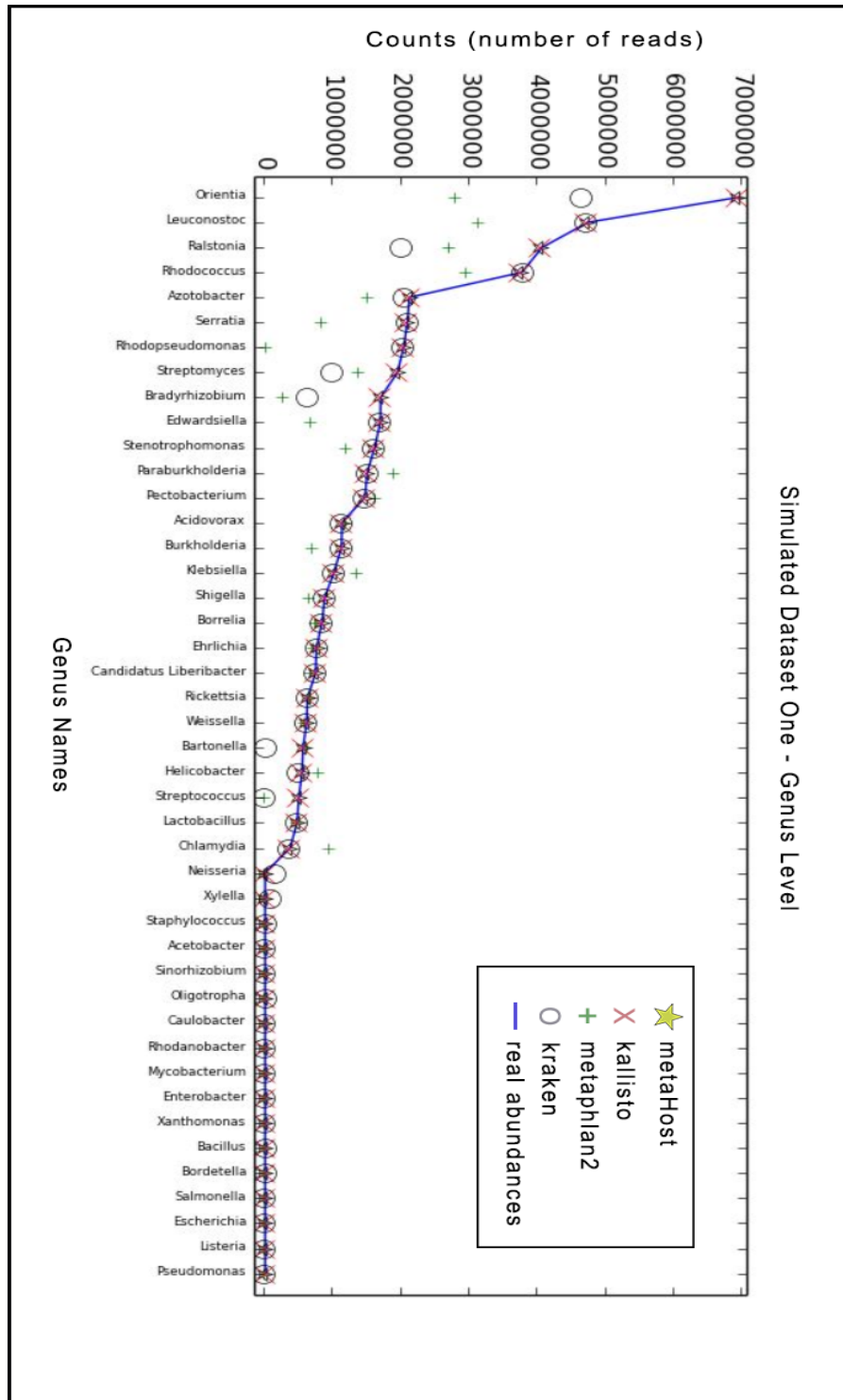
Σχήμα 12 - metaHost Dataset 5 - Species level



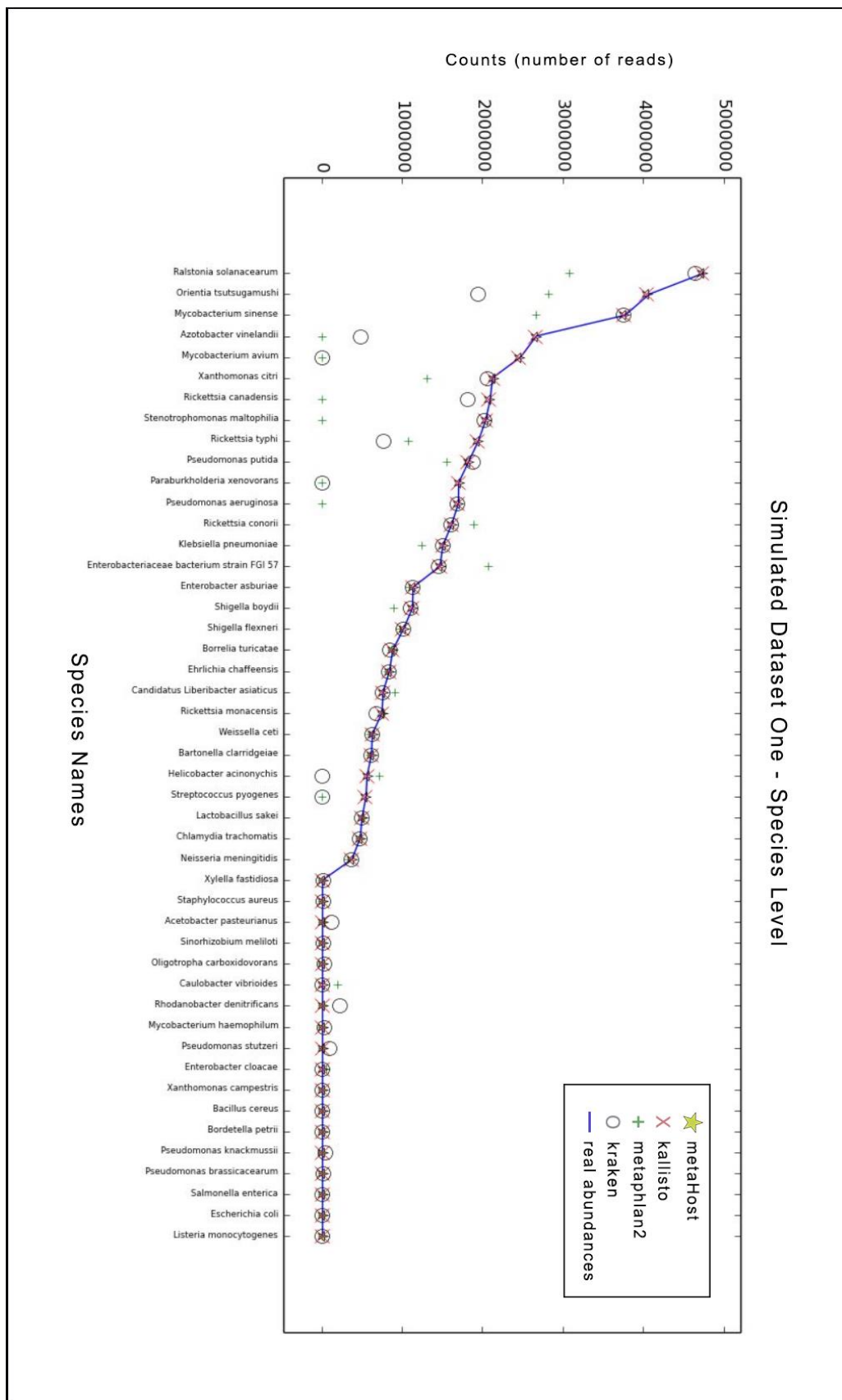
Σχήμα 13 - metaHost Dataset 5 - Strain level

3.3 Σύγκριση του metaHost – Γραφήματα

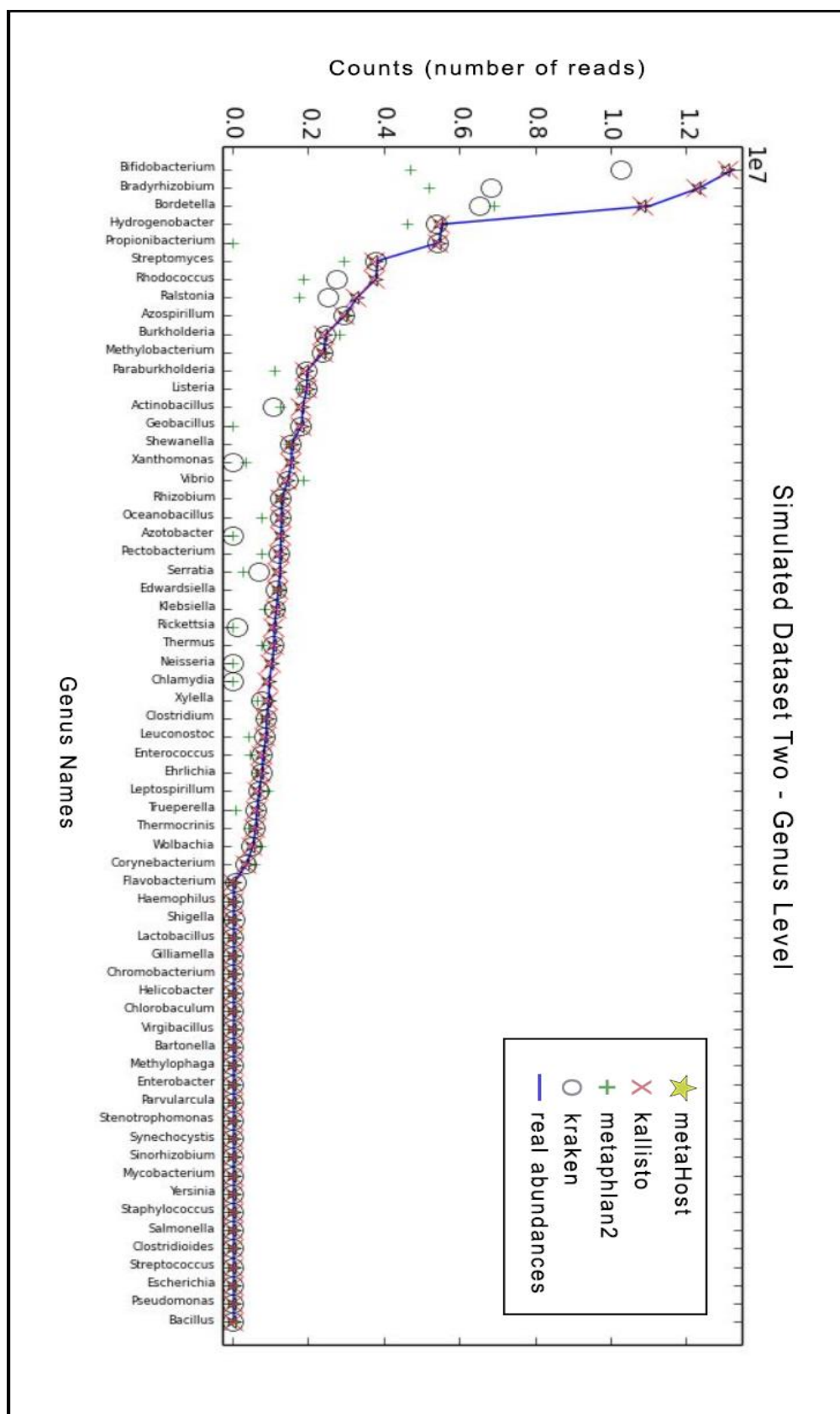
Η σύγκριση του metaHost, έγινε με το MetaPhlAn2, το Kraken και το Kallisto.



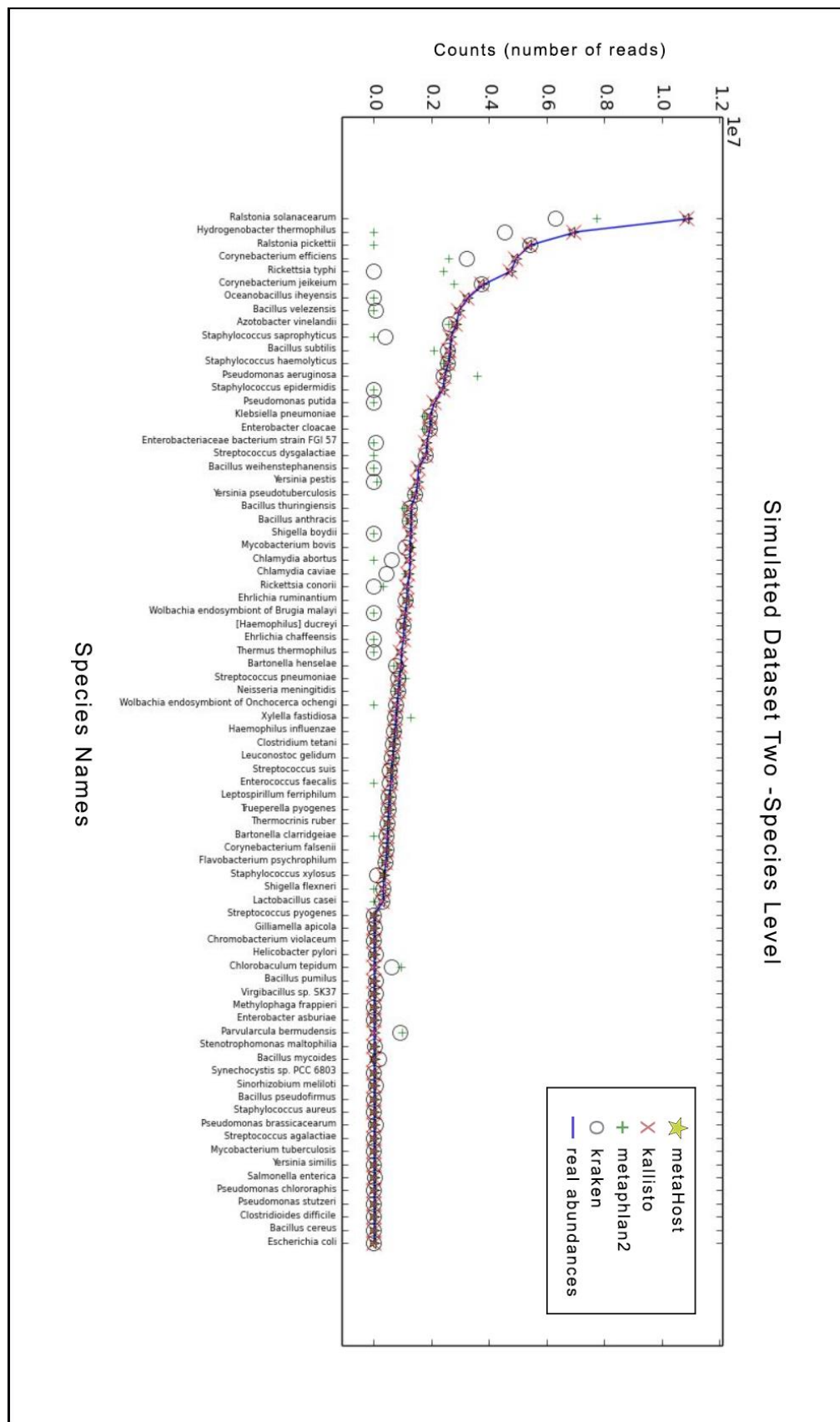
Σχήμα 14 - Σύγκριση αλγόριθμων Dataset 1 - Genus level



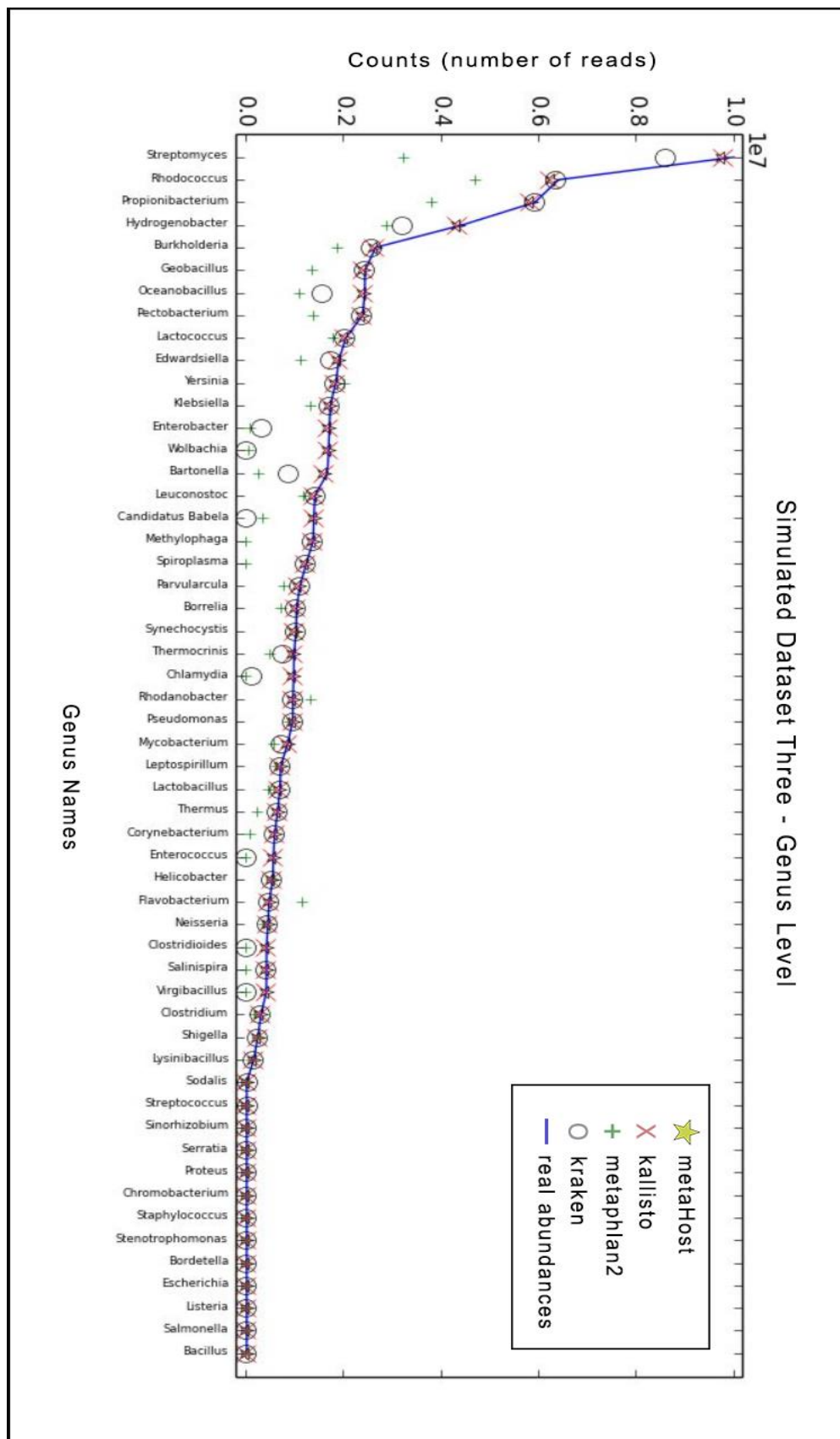
Σχήμα 15 - Σύγκριση αλγόριθμων Dataset 1 - Species level



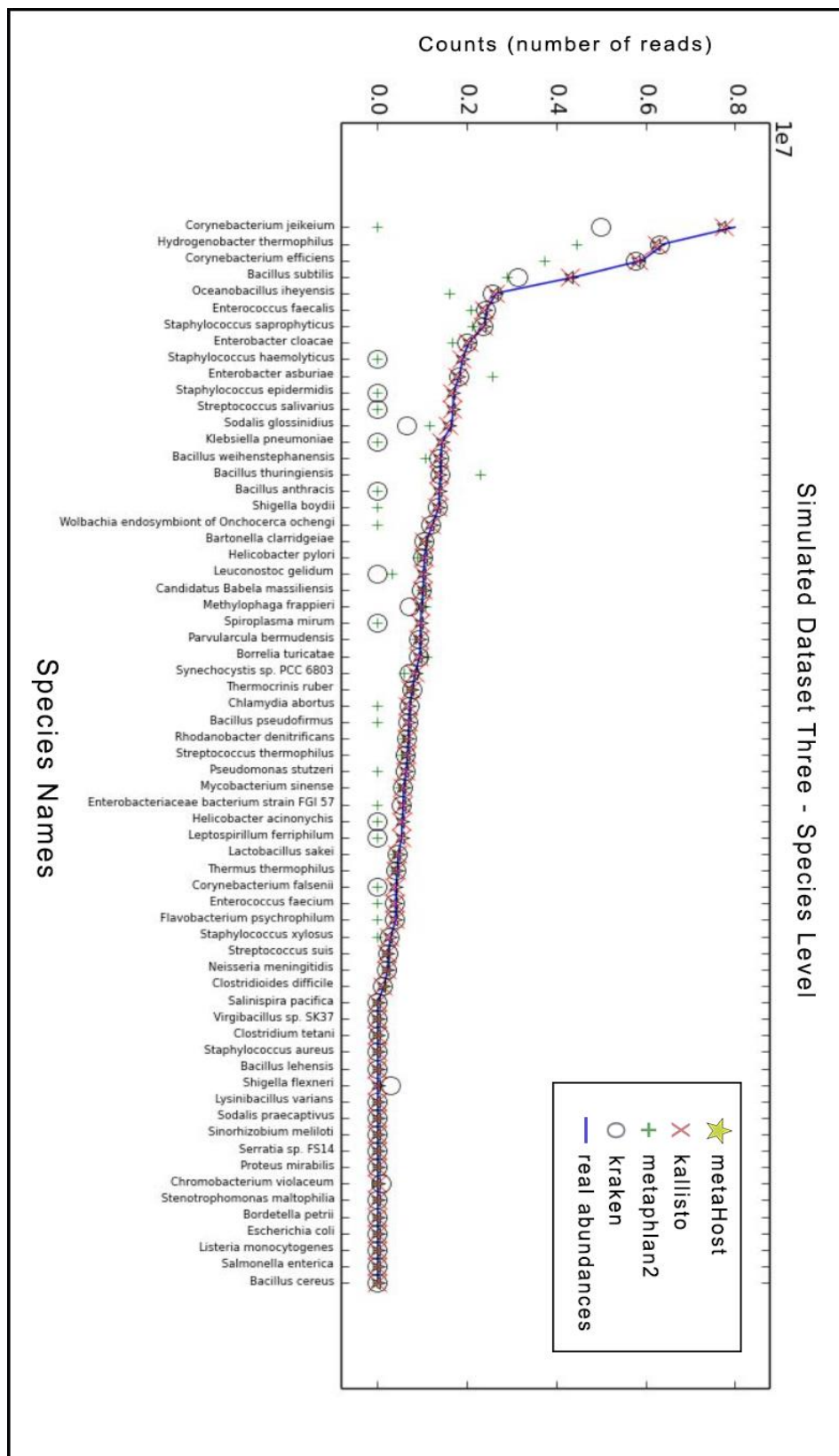
Σχήμα 16 - Σύγκριση αλγόριθμων Dataset 2 - Genus level



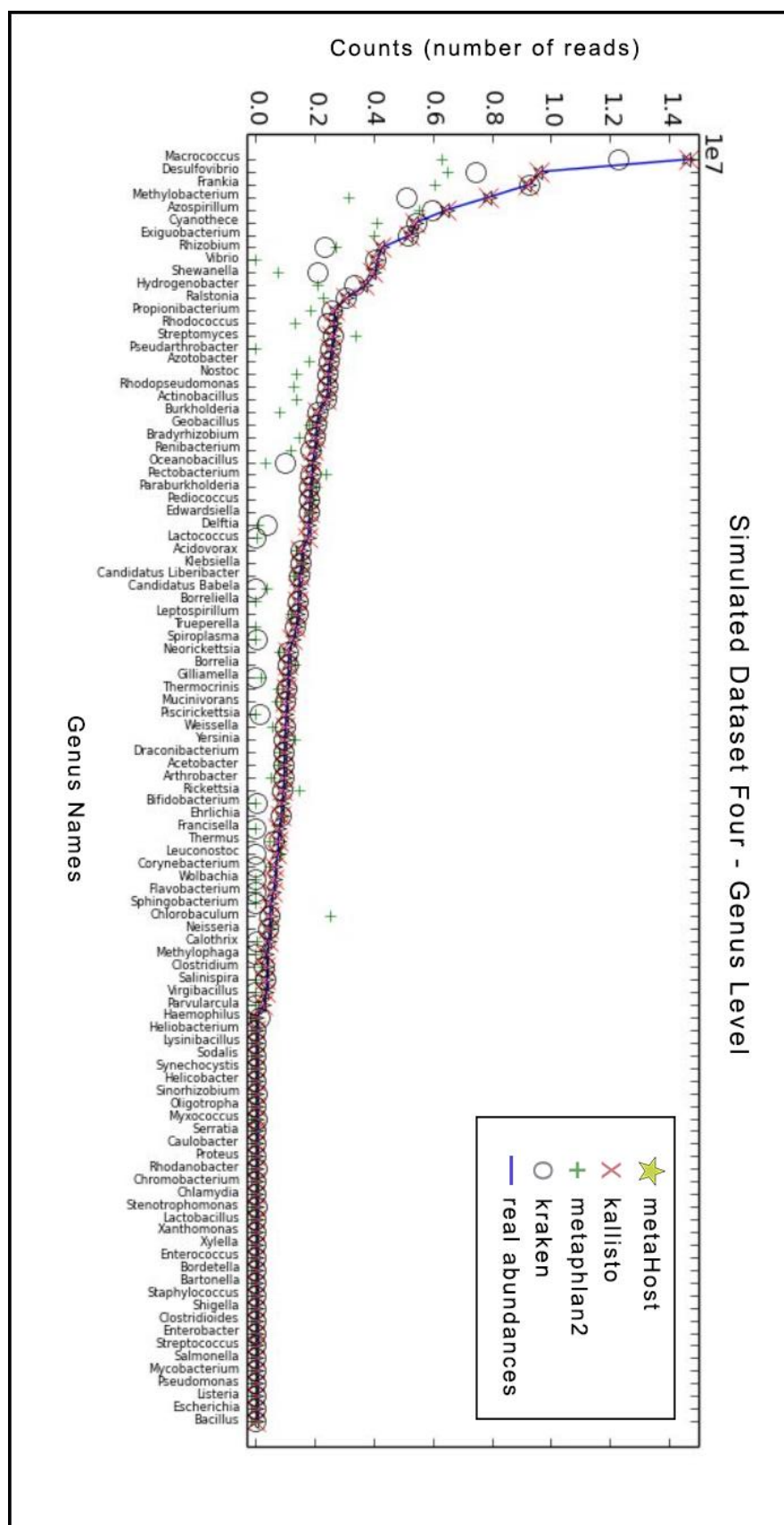
Σχήμα 17 - Σύγκριση αλγόριθμων Dataset 2 - Species level



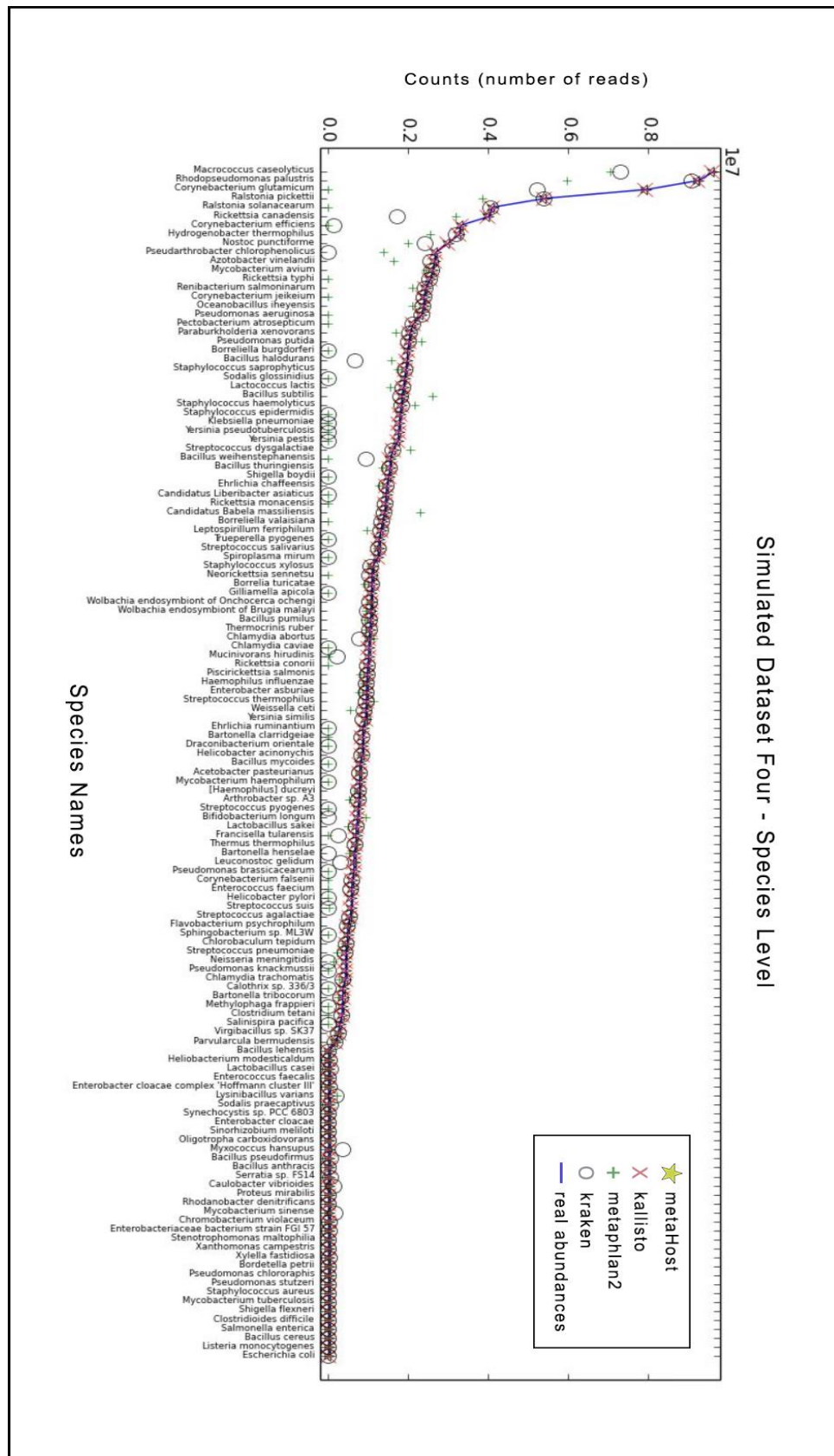
Σχήμα 18 - Σύγκριση αλγόριθμων Dataset 3 - Genus level



Σχήμα 19 - Σύγκριση αλγορίθμων Dataset 3 - Species level

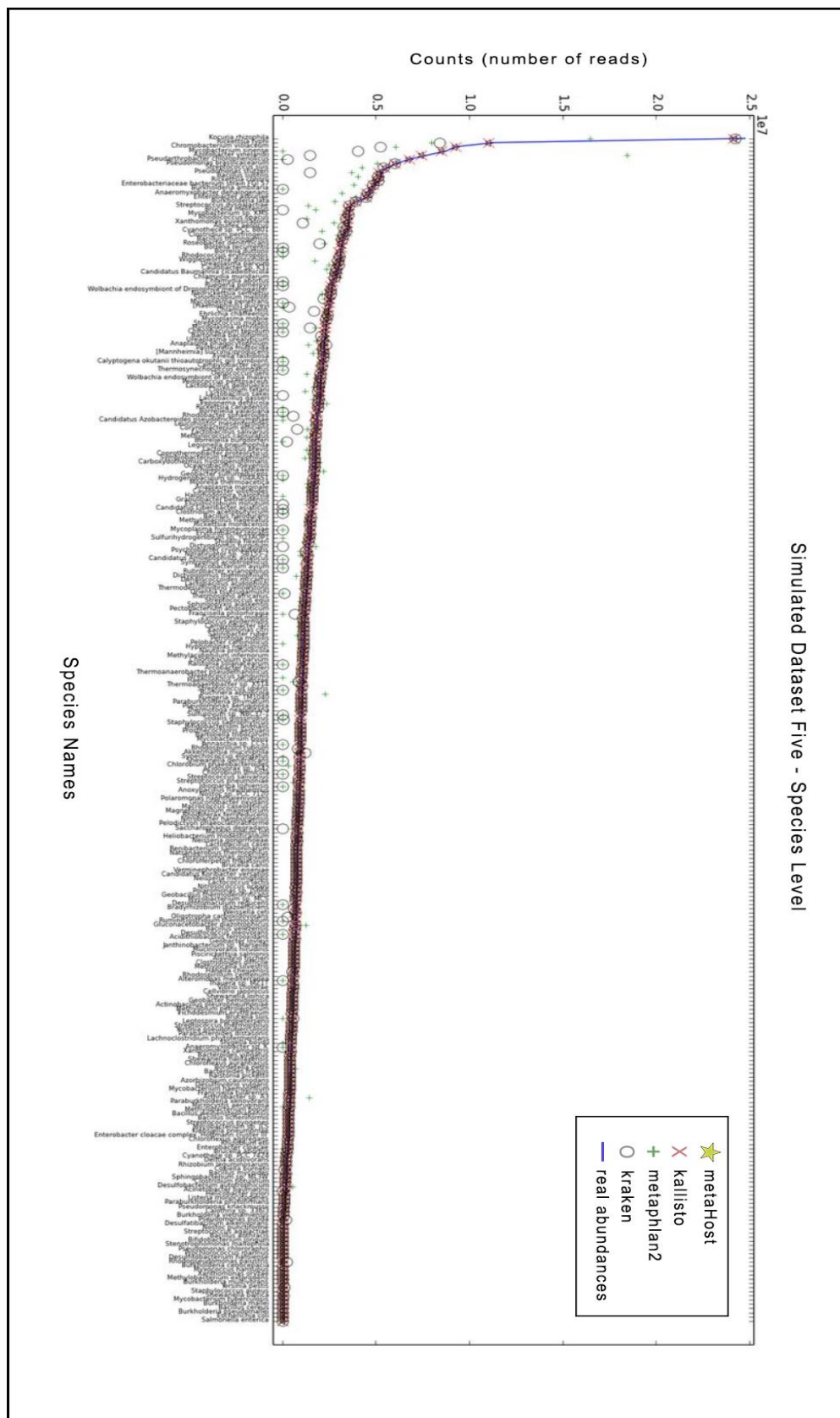


Σχήμα 20 - Σύγκριση αλγόριθμων Dataset 4 - Genus level



Σχήμα 21 - Σύγκριση αλγόριθμων Dataset 4 - Species level





Σχήμα 23 - Σύγκριση αλγόριθμων Dataset 5 - Species level

3.4 Σύγκριση του metaHost – Μετρικές (metrics)

Εκτός της οπτικοποίησης των αποτελεσμάτων μέσω γραφικών παραστάσεων, χρησιμοποιήθηκαν και κάποιες μετρικές με σκοπό την ακριβέστερη απόδοση των αποτελεσμάτων των συγκρίσεων μεταξύ των προγραμμάτων.

Οι πρώτες μετρικές που επιλέχθηκαν, είναι μετρικές συσχέτισης όπου ουσιαστικά μας δείχνουν πόσο συσχετισμένα είναι τα πραγματικά δεδομένα σε σχέση με τα αποτελέσματα των προγραμμάτων. Όσο πιο κοντά στην μονάδα είναι οι εν λόγω μετρικές, τόσο μεγαλύτερη συσχέτιση υπάρχει και άρα τόσο καλύτερα τα αποτελέσματα. Οι μετρικές συσχέτισης που χρησιμοποιήθηκαν είναι οι Pearson Correlation Coefficient, Spearman Rho και Kendall's Tau.

Η επόμενη κατηγορία από μετρικές που επιλέχθηκε, είναι οι μετρικές ομοιότητας όπου μας παρουσιάζουν πόσο όμοια είναι τα πραγματικά δεδομένα με τα αποτελέσματα. Όσο πιο κοντά βρίσκεται αυτό το είδος από μετρικές στη μονάδα, τόσο πιο όμοια τα πραγματικά δεδομένα με τα αποτελέσματα και άρα τόσο μεγαλύτερη η ακρίβεια αυτών. Οι μετρικές ομοιότητας που επιλέχθηκαν είναι οι Cosine Similarity και Morisita Horn.

Στην συνέχεια επιλέχθηκαν δύο μετρικές απόστασης όπου ουσιαστικά υπολογίζουν τις αποστάσεις των πραγματικών τιμών από τα αποτελέσματα των προγραμμάτων σαν απόλυτες τιμές. Προφανώς, όσο μικρότερες είναι οι αποστάσεις, τόσο πιο ακριβή τα αποτελέσματα. Οι μετρικές αποστάσεως που επιλέχθηκαν είναι οι Euclidean Distance και Manhattan Distance.

Τέλος, επιλέχθηκε και μια μετρική ανομοιότητας όπου μας παρουσιάζει πόσο ανόμοια είναι τα πραγματικά δεδομένα από τα αποτελέσματα των προγραμμάτων. Εδώ, ισχύει το ακριβώς αντίθετο από τις μετρικές ομοιότητας, δηλαδή όσο πιο κοντά είναι η μετρική στη μονάδα τόσο πιο ανόμοια τα αποτελέσματα. Αυτό σημαίνει πως όσο πιο κοντά είναι η μετρική στο μηδέν, τόσο πιο όμοια τα αποτελέσματα και άρα τόσο πιο ακριβή.

Παρακάτω, ακολουθούν αναλυτικοί πίνακες με όλες τις μετρικές που αναφέρθηκαν παραπάνω, για όλα τα συνθετικά σύνολα δεδομένων και όλα τα προγράμματα που συγκρίθηκαν.

Πίνακας 2 - Metrics dataset 1 - Genus level

Genus Level	Simulated dataset one			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.205	0.216	0.172	0.462
Pearson Correlation Coefficient	0.999	0.999	0.874	0.948
Spearman Rho	0.973	0.965	0.889	0.916
Kendall's Tau	0.939	0.926	0.750	0.808
Cosine Similarity	0.999	0.999	0.919	0.966
Minkowski distance (Euclidean)	144736	121763	18502719	7917837
Minkowski distance (Manhattan)	47398	38451	5755944	3486362
Bray Curtis	0.001	0.001	0.241	0.093
Morisita Horn	0.999	0.999	0.916	0.965

Πίνακας 3 - Metrics dataset 1 - Species level

Species Level	Simulated dataset one			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.234	0.246	0.255	0.503
Pearson Correlation Coefficient	0.999	0.999	0.742	0.844
Spearman Rho	0.973	0.971	0.587	0.695
Kendall's Tau	0.935	0.928	0.569	0.624
Cosine Similarity	0.999	0.999	0.837	0.903
Minkowski distance (Euclidean)	144970	104817	21258940	11996323
Minkowski distance (Manhattan)	44832	27712	5940212	4509694
Bray Curtis	0.001	0.002	0.290	0.148
Morisita Horn	0.999	0.999	0.833	0.901

Πίνακας 4 - Metrics dataset 2 - Genus level

Genus Level	Simulated dataset two			
	Salmon	Kallisto	MetaPhlAn2	Kraken
MARD	0.314	0.314	0.221	0.491
Pearson Correlation Coefficient	0.999	0.999	0.872	0.960
Spearman Rho	0.969	0.967	0.833	0.819
Kendall's Tau	0.921	0.921	0.730	0.733
Cosine Similarity	0.999	0.999	0.905	0.970
Minkowski distance (Euclidean)	667839	654734	45390498	22294440
Minkowski distance (Manhattan)	239925	229826	13801549	8269972
Bray Curtis	0.003	0.003	0.295	0.128
Morisita Horn	0.999	0.999	0.903	0.968

Πίνακας 5 - Metrics dataset 2 - Species level

Species Level	Simulated dataset two			
	Salmon	Kallisto	MetaPhlAn2	Kraken
MARD	0.305	0.305	0.343	0.530
Pearson Correlation Coefficient	0.999	0.999	0.686	0.816
Spearman Rho	0.981	0.971	0.471	0.444
Kendall's Tau	0.941	0.933	0.424	0.427
Cosine Similarity	0.999	0.999	0.768	0.870
Minkowski distance (Euclidean)	701641	688502	56703221	41410397
Minkowski distance (Manhattan)	229269	219094	12701544	10181468
Bray Curtis	0.003	0.003	0.385	0.261
Morisita Horn	0.999	0.999	0.756	0.867

Πίνακας 6 - Metrics dataset 3 - Genus level

Genus Level	Simulated dataset three			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.187	0.194	0.300	0.379
Pearson Correlation Coefficient	0.999	0.999	0.863	0.970
Spearman Rho	0.993	0.992	0.810	0.817
Kendall's Tau	0.971	0.971	0.679	0.736
Cosine Similarity	0.999	0.999	0.908	0.978
Minkowski distance (Euclidean)	840604	821891	32897332	11666891
Minkowski distance (Manhattan)	321665	312981	8690294	3579986
Bray Curtis	0.006	0.006	0.307	0.092
Morisita Horn	0.999	0.999	0.908	0.976

Πίνακας 7 - Metrics dataset 3 - Species level

Species Level	Simulated dataset three			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.264	0.264	0.356	0.448
Pearson Correlation Coefficient	0.999	0.999	0.668	0.920
Spearman Rho	0.989	0.983	0.594	0.577
Kendall's Tau	0.960	0.956	0.527	0.551
Cosine Similarity	0.999	0.999	0.764	0.943
Minkowski distance (Euclidean)	861921	851721	36755916	18196670
Minkowski distance (Manhattan)	322527	311366	9916390	5248428
Bray Curtis	0.006	0.006	0.346	0.149
Morisita Horn	0.999	0.999	0.758	0.939

Πίνακας 8 - Metrics dataset 4 - Genus level

Genus Level	Simulated dataset four			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.119	0.127	0.314	0.504
Pearson Correlation Coefficient	0.999	0.999	0.878	0.971
Spearman Rho	0.986	0.984	0.820	0.827
Kendall's Tau	0.966	0.962	0.684	0.705
Cosine Similarity	0.999	0.999	0.913	0.979
Minkowski distance (Euclidean)	340909	324827	70767396	27906521
Minkowski distance (Manhattan)	326526	302837	13816189	6427463
Bray Curtis	0.001	0.001	0.292	0.101
Morisita Horn	0.999	0.999	0.913	0.978

Πίνακας 9 - Metrics dataset 4 - Species level

Species Level	Simulated dataset four			
	metaHost	Kallisto	MetaPhlAn2	Kraken
MARD	0.165	0.165	0.389	0.514
Pearson Correlation Coefficient	0.999	0.999	0.735	0.900
Spearman Rho	0.991	0.991	0.499	0.539
Kendall's Tau	0.969	0.969	0.446	0.469
Cosine Similarity	0.999	0.999	0.805	0.926
Minkowski distance (Euclidean)	350643	340891	82524507	47382861
Minkowski distance (Manhattan)	94888	92811	13649887	8619547
Bray Curtis	0.001	0.001	0.355	0.180
Morisita Horn	0.999	0.999	0.786	0.916

Πίνακας 10 - Metrics dataset 5 - Genus level

Genus Level	Simulated dataset five			
	Salmon	Kallisto	MetaPhlAn2	Kraken
MARD	0.049	0.054	0.243	0.136
Pearson Correlation Coefficient	0.999	0.999	0.927	0.987
Spearman Rho	0.999	0.997	0.730	0.897
Kendall's Tau	0.996	0.995	0.614	0.874
Cosine Similarity	0.999	0.999	0.946	0.990
Minkowski distance (Euclidean)	3457201	3467002	157187958	33172024
Minkowski distance (Manhattan)	944360	944578	25312964	8466418
Bray Curtis	0.004	0.004	0.220	0.041
Morisita Horn	0.999	0.999	0.943	0.989

Πίνακας 11 - Metrics dataset 5 - Species level

Species Level	Simulated dataset five			
	Salmon	Kallisto	MetaPhlAn2	Kraken
MARD	0.066	0.066	0.306	0.232
Pearson Correlation Coefficient	0.999	0.999	0.786	0.886
Spearman Rho	0.999	0.999	0.583	0.636
Kendall's Tau	0.993	0.993	0.538	0.626
Cosine Similarity	0.999	0.999	0.842	0.918
Minkowski distance (Euclidean)	3805061	3903362	170893284	104451950
Minkowski distance (Manhattan)	781619	809103	22869963	17036557
Bray Curtis	0.004	0.004	0.245	0.143
Morisita Horn	0.999	0.999	0.830	0.912

3.5 Χρόνοι εκτέλεσης

Πίνακας 12 - Χρόνοι εκτέλεσης αλγόριθμων

Time – 24 cores					
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
metaHost	5m49.407s	8m30.703s	7m39.853s	8m17.569s	27m9.700s
MetaPhlAn2	4m31.347s	7m19.760s	6m24.280s	13m56.735s	36m6.477s
Kraken	7m12s	13m12.778s	10m35.785s	19m56.253s	64m37.912s

4. ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

4.1 Τα δεδομένα

Τα δεδομένα που αναλύθηκαν προέρχονται από τον οργανισμό *Mus Musculus* και από το πείραμα RNA-Seq. Τα δείγματα είναι όλα ιστοειδικά, κυρίως από περιοχές του εντέρου, εκτός από ένα από αυτά που είναι δείγμα περιττωμάτων με σκοπό την χαρτογράφηση του μικροβιώματος με το πείραμα 16S rRNA-Seq.

Πίνακας 13 - Πληροφορίες των δεδομένων που αναλύθηκαν

BioProjects ID	Library Prep	Instrument Model	Source Name	Layout	Num of Samples	AVG number of reads
PRJNA246715	RNA-Seq	HiSeq 2000	Colon epithel	PE	10	38M
PRJNA279201	RNA-Seq	HiSeq 2000	Colon polyp	PE	20	36M
PRJNA193402	RNA-Seq	HiSeq 2000	Ileum, colon, caecum	PE	6	40M
PRJNA169535	RNA-Seq	Analyzer IIx	Intestine, ileum, adenomas	SE	11	14M
PRJNA285523	16S rRNA-Seq	MiSeq	Mucosal Sample (intestinal microbiota)	SE	30	36K
PRJNA248400	RNA-Seq	HiSeq 2000	Intestine mucosa	SE	9	160M
PRJNA257651	RNA-Seq	HiSeq 2000	small intestinal tissue (ileum)	SE	9	44M
PRJNA254957	RNA-Seq	HiSeq 2000	small intestinal crypts, adenomas	PE	18	35M
PRJNA177791	RNA-Seq	HiSeq 2000	colon	PE	3	75M
GSM900198 (GEO)	RNA-Seq	Analyzer II	Colon (ENCODE)	PE	6	21M
GSM900189 (GEO)	RNA-Seq	Analyzer II	Large intestine (ENCODE)	PE	4	37M
GSM900186 (GEO)	RNA-Seq	Analyzer II	Small intestine (ENCODE)	PE	10	15M

4.2 Η βάση των μικροβίων (reference)

Για τη δημιουργία της βάσης των μικροβίων, μελετήθηκαν σχετικές εργασίες οι οποίες παρουσιάζουν μικροοργανισμούς που βρέθηκαν σε NGS δείγματα (16S rRNA Sequencing, Shotgun Sequencing) από ποντίκια (Mouse - *Mus musculus*). Το μεγαλύτερο μέρος των γονιδιωμάτων των βακτηρίων που περιέχονται στο reference, προέρχονται από την εργασία [45] καθώς επίσης και τις εργασίες [46], [47], [48]. Ο αριθμός των γονιδιωμάτων που παρουσιάζονται στις παραπάνω εργασίες είναι σχετικά μικρός, για το λόγο αυτό, προστέθηκαν στο reference και τα υπόλοιπα γονιδιώματα των ειδών στα οποία ανήκουν. Σχετικά με τους ιούς, λόγω του μικρούς μεγέθους των γονιδιωμάτων τους, προστέθηκαν όλα όσα υπάρχουν στη βάση του NCBI. Τέλος, προστέθηκαν και όλα τα γονιδιώματα των Αρχαίων (Archaea).

Όλα τα γονιδιώματα που περιέχονται στη βάση, προέρχονται από τη το NCBI (refseq) και είναι ολοκληρωμένα (Completed genomes) και ανανεωμένα με βάση την τελευταία έκδοση (latest).

Τελικώς, ο συνολικός αριθμός των βακτηρίων ανέρχεται σε 2.100 sequences (γονιδιώματα, πλασμίδια), ο αριθμός των ιών σε 6.500 γονιδιώματα και ο αριθμός των αρχαίων σε 560. Το μέγεθος της βάσης είναι 5.0 GB.

4.3 Αποτελέσματα

Για τον υπολογισμό των αποτελεσμάτων αρχικά έγινε φιλτράρισμα ώστε να παραμείνουν μόνο όσα γονιδιώματα είχαν περισσότερα από 50 reads. Στη συνέχεια βρέθηκε ο μέσος όρος των reads για κάθε σύνολο δεδομένων (datasets) και τέλος έγινε πάλι φιλτράρισμα κρατώντας μόνο τους μέσους όρους που ήταν μεγαλύτεροι των 80 reads. Ο μέσος όρος των reads υπολογίστηκε προσθέτοντας τα reads όλων των δειγμάτων (samples) του κάθε συνόλου δεδομένων και στη συνέχεια διαιρώντας με τον αριθμό των δειγμάτων. Με αυτή τη διαδικασία διασφαλίστηκε πως στα αποτελέσματα θα παραμείνουν όλα τα μικρόβια (strains) με μεγάλη και μεσαία αφθονία, ενώ αυτά με πολύ μικρή αφθονία θα απομακρυνθούν. Παρακάτω παρατίθενται τα 15 γονιδιώματα με τη μεγαλύτερη αφθονία για κάθε σύνολο δεδομένων ξεχωριστά.

Σε κάποιες περιπτώσεις ο αριθμός των γονιδιωμάτων που παρουσιάζονται είναι μικρότερος από 15, αυτό οφείλεται στο γεγονός πως τα μικρόβια που βρέθηκαν στο δείγμα ήταν λιγότερα σε αριθμό.

Στους πίνακες που ακολουθούν, στο πρώτο πεδίο παρατηρούμε τον κωδικό καθενός από τα μικρόβια όπως αυτός έχει αποδοθεί στο NCBI, στο δεύτερο πεδίο το όνομα του μικροβίου και τέλος στο τρίτο πεδίο τον αριθμό των reads (counts) του μικροβίου που βρέθηκαν στο δείγμα.

Πίνακας 14 - Αποτελέσματα dataset PRJNA246715

Dataset PRJNA246715		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	5067
NZ_CP014577.1	Salmonella enterica subsp. enterica serovar Typhimurium strain RM9437	1294
NZ_LN515531.1	Methanobacterium formicicum genome assembly DSM1535	1174
NC_008168.1	Choristoneura occidentalis granulovirus	784
NC_001702.1	Murine type C retrovirus	619
NC_022098.1	Pandoravirus salinus	500
NC_001416.1	Enterobacteria phage lambda	349
NC_020235.1	Rosellinia necatrix partitivirus 2	277
NC_001506.1	Murine osteosarcoma virus	254
NC_004102.1	Hepatitis C virus genotype 1	182
NC_007609.1	Dulcamara mottle virus	177
NC_021858.1	Pandoravirus dulcis	174
NC_009668.1	Ochrobactrum anthropi ATCC 49188	152
NC_009720.1	Xanthobacter autotrophicus Py2	136

Πίνακας 15 - Αποτελέσματα dataset PRJNA279201

Dataset PRJNA279201		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	21053
NC_006273.2	Human herpesvirus 5 strain Merlin	6850
NC_001669.1	Simian virus 40	2096
NC_001702.1	Murine type C retrovirus	1623
NZ_CP015725.1	Salmonella enterica strain C629	1225
NZ_CP012141.1	Shigella flexneri 4c strain 120	1055
NC_014976.1	Bacillus subtilis BSn5	721
NC_007815.2	PreXMRV-1 provirus	525
NC_001500.1	Spleen focus-forming virus	449
NZ_CP011335.1	Escherichia coli O104:H4 str. C227-11	420
NZ_CP015837.1	Escherichia coli strain MS6198	353
NZ_CP015078.1	Escherichia coli strain EcoI_448	326
NC_003287.2	Enterobacteria phage M13	317
NZ_CP007392.1	Escherichia coli strain ST2747	312

Πίνακας 16 - Αποτελέσματα dataset PRJNA193402

Dataset PRJNA193402		
Code	Strain Name	AVG Counts
NZ_LN515531.1	Methanobacterium formicicum genome assembly DSM1535	60758
NC_008168.1	Choristoneura occidentalis granulovirus	4803
NZ_CP014577.1	Salmonella enterica subsp. enterica serovar Typhimurium strain RM9437	1455
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	1330
NC_022098.1	Pandoravirus salinus	1033
NC_017618.1	Streptococcus suis JS14	536
NC_021858.1	Pandoravirus dulcis	387
NC_017621.1	Streptococcus suis D12	346
NC_020235.1	Rosellinia necatrix partitivirus 2	297
NC_007609.1	Dulcamara mottle virus	276
NC_004812.1	Macacine herpesvirus 1	269
NC_004102.1	Hepatitis C virus genotype 1	251
NC_008912.1	Glypta fumiferanae ichnovirus segment C9	183
NC_001493.2	Ictalurid herpesvirus 1 strain Auburn 1	177

Πίνακας 17 - Αποτελέσματα dataset PRJNA169535

Dataset PRJNA169535		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	3604
NC_018476.1	Simbu virus RdRp gene for RNA-dependent RNA polymerase	3080
NC_004917.1	Helicobacter hepaticus ATCC 51449	2359
NC_028250.1	Rosellinia necatrix partitivirus 6	2044
NC_004122.1	Spring beauty latent virus RNA 3	1255
NC_028095.1	Torulaspora delbrueckii dsRNA Mbarr-1 killer virus strain EX1180	1191
NC_023685.1	Rhizoctonia solani dsRNA virus 2	1069
NC_021095.1	White clover cryptic virus 2 isolate IPP_Lirepa segment RNA 2	997
NC_021097.1	Red clover cryptic virus 2 isolate IPP_Nemaro segment RNA 2	975
NC_003542.1	Cowpea chlorotic mottle virus RNA 3	626
NC_001782.1	Saccharomyces cerevisiae killer virus M1	622
NC_007609.1	Dulcamara mottle virus	603
NC_004102.1	Hepatitis C virus genotype 1	586
NC_024502.1	Gentian ovary ring-spot virus genomic RNA	410
NC_021148.1	Dill cryptic virus 2 isolate IPP_hortorum	404

Πίνακας 18 - Αποτελέσματα dataset PRJNA248400

Dataset PRJNA248400		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	140100
NC_001702.1	Murine type C retrovirus	20759
NC_008912.1	Glypta fumiferanae ichtnovirus segment C9	11826
NZ_CP014348.1	Escherichia coli str. K-12 substr. MG1655 strain JW5437-1	11609
NC_022098.1	Pandoravirus salinus	10131
NC_030925.1	Bacillus phage Shbh1	9525
NC_007815.2	PreXMRV-1 provirus	8504
NZ_CP010439.1	Escherichia coli K-12 strain ER3440	7604
NC_001493.2	Ictalurid herpesvirus 1 strain Auburn 1	6458
NZ_CP009515.1	Methanosarcina lacustris Z-7289	5853
NC_001500.1	Spleen focus-forming virus	5233
NZ_CP015159.1	Escherichia coli strain Eco889	4656
NC_008913.1	Glypta fumiferanae ichtnovirus segment C10	4475
NC_009127.1	Cyprinid herpesvirus 3	4363

Πίνακας 19 - Αποτελέσματα dataset PRJNA257651

Dataset PRJNA257651		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	132

Πίνακας 20 - Αποτελέσματα dataset PRJNA254957

Dataset PRJNA254957		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	13191
NC_001702.1	Murine type C retrovirus	1417
NC_017618.1	Streptococcus suis JS14	467
NC_001500.1	Spleen focus-forming virus	369
NC_007815.2	PreXMRV-1 provirus	283
NC_001501.1	Moloney murine leukemia virus	112
NC_022098.1	Pandoravirus salinus	96
NC_001362.1	Friend murine leukemia virus FB29	84
NC_017621.1	Streptococcus suis D12	82
NZ_CP014577.1	Salmonella enterica subsp. enterica serovar Typhimurium strain RM9437	81

Πίνακας 21 - Αποτελέσματα dataset PRJNA177791

Dataset PRJNA177791		
Code	Strain Name	AVG Counts
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	56421
NC_008168.1	Choristoneura occidentalis granulovirus	15307
NC_001702.1	Murine type C retrovirus	7742
NZ_CP014348.1	Escherichia coli str, K-12 substr, MG1655 strain JW5437-1	4231
NC_030925.1	Bacillus phage Shbh1	3796
NC_009127.1	Cyprinid herpesvirus 3	2109
NC_004102.1	Hepatitis C virus genotype 1	1757
NC_007815.2	PreXMRV-1 provirus	1468
NZ_CP011113.1	Escherichia coli strain RR1	1064
NC_007609.1	Dulcamara mottle virus	1031
NZ_LN832404.1	Escherichia coli K-12 genome assembly EcoliK12AG100	982
NZ_CP009505.1	Methanosarcina sp, MTP4	873
NZ_CP015159.1	Escherichia coli strain Eco889	871
NC_022098.1	Pandoravirus salinus	835

Πίνακας 22 - Αποτελέσματα dataset GSM900198 rep1

Dataset GSM900198 ENCODE – rep1		
Code	Strain Name	AVG Counts
NZ_CP015975.1	Bacillus subtilis subsp. subtilis strain delta6	38673
NC_012781.1	Eubacterium rectale ATCC 33656	22104
NC_014388.1	Butyrivibrio proteoclasticus B316	15996
NZ_LN515531.1	Methanobacterium formicicum genome assembly DSM1535	10165
NC_010001.1	Clostridium phytofermentans ISDg	7052
NC_009656.1	Pseudomonas aeruginosa PA7	3757
NZ_CP013993.1	Pseudomonas aeruginosa DHS01	2602
NC_014833.1	Ruminococcus albus 7	2350
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	2126
NC_008168.1	Choristoneura occidentalis granulovirus	1991
NZ_CP008857.1	Pseudomonas aeruginosa strain F30658	1687
NC_015164.1	Bacteroides salanitronis DSM 18170	1449
NZ_CP008866.1	Pseudomonas aeruginosa strain T38079	1322
NZ_AP012496.1	Bacillus subtilis BEST7003	1141
NC_023149.1	Pseudomonas aeruginosa SCV20265	1054

Πίνακας 23 - Αποτελέσματα dataset GSM900198 - rep2

Dataset GSM900198 ENCODE – rep2		
Code	Strain Name	AVG Counts
NZ_CP015975.1	Bacillus subtilis subsp, subtilis strain delta6	43596
NC_012781.1	Eubacterium rectale ATCC 33656	14044
NC_014388.1	Butyrivibrio proteoclasticus B316	11155
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	7183
NC_010001.1	Clostridium phytofermentans ISDg	6081
NC_015164.1	Bacteroides salanitronis DSM 18170	4427
NC_008168.1	Choristoneura occidentalis granulovirus	4374
NC_009615.1	Parabacteroides distasonis ATCC 8503	3746
NC_009656.1	Pseudomonas aeruginosa PA7	2759
NZ_CP013993.1	Pseudomonas aeruginosa DHS01	2119
NC_014833.1	Ruminococcus albus 7	2097
NC_018011.1	Alistipes finegoldii DSM 17242	1903
NC_014033.1	Prevotella ruminicola 23	1892
NC_014933.1	Bacteroides helcogenes P 36-108	1583
NZ_AP012496.1	Bacillus subtilis BEST7003	1484

Πίνακας 24 - Αποτελέσματα dataset GSM900189 rep2

Dataset GSM900189 ENCODE – rep2		
Code	Strain Name	AVG Counts
NZ_CP015975.1	Bacillus subtilis subsp, subtilis strain delta6	580379
NC_000909.1	Methanocaldococcus jannaschii DSM 2661	10794
NC_002167.1	Enterobacteria phage HK97	9694
NZ_CP015375.1	Bacillus subtilis subsp, subtilis strain KCTC 3135	8497
NC_008168.1	Choristoneura occidentalis granulovirus	8207
NZ_AP012496.1	Bacillus subtilis BEST7003	7446
NZ_LN515531.1	Methanobacterium formicum genome assembly DSM1535	7164
NC_009656.1	Pseudomonas aeruginosa PA7	7080
NZ_CP013993.1	Pseudomonas aeruginosa DHS01	6159
NC_017621.1	Streptococcus suis D12	2435
NC_023149.1	Pseudomonas aeruginosa SCV20265	2433
NC_014976.1	Bacillus subtilis BSn5	1994
NZ_CP008861.1	Pseudomonas aeruginosa strain H47921	1909
NC_008435.1	Rhodopseudomonas palustris BisA53	1764
NC_010943.1	Stenotrophomonas maltophilia K279a	1674

Πίνακας 25 - Αποτελέσματα dataset GSM900186 rep1

Dataset GSM900186 ENCODE – rep1		
Code	Strain Name	AVG Counts
NZ_CP015975.1	Bacillus subtilis subsp, subtilis strain delta6	13962
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	1369
NC_008168.1	Choristoneura occidentalis granulovirus	1202
NC_009656.1	Pseudomonas aeruginosa PA7	924
NC_017621.1	Streptococcus suis D12	616
NZ_AP012496.1	Bacillus subtilis BEST7003	511
NC_017618.1	Streptococcus suis JS14	457
NC_022098.1	Pandoravirus salinus	422
NC_008912.1	Glypta fumiferanae ichnovirus segment C9	385
NZ_CP008870.1	Pseudomonas aeruginosa strain W36662	352
NC_018476.1	Simbu virus RdRp gene for RNA-dependent RNA polymerase	347
NZ_CP015375.1	Bacillus subtilis subsp, subtilis strain KCTC 3135	347
NC_019491.1	Cyprinid herpesvirus 1 strain NG-J1	281
NC_002167.1	Enterobacteria phage HK97	229
NC_000909.1	Methanocaldococcus jannaschii DSM 2661	182

Πίνακας 26 - Αποτελέσματα dataset GSM900186 rep2

Dataset GSM900186 ENCODE – rep2		
Code	Strain Name	AVG Counts
NZ_CP015975.1	Bacillus subtilis subsp, subtilis strain delta6	19626
NC_029853.1	Mus musculus mobilized endogenous polytropic provirus clone 15	2320
NC_008168.1	Choristoneura occidentalis granulovirus	1249
NC_009656.1	Pseudomonas aeruginosa PA7	1198
NC_022098.1	Pandoravirus salinus	984
NC_017621.1	Streptococcus suis D12	904
NC_008912.1	Glypta fumiferanae ichnovirus segment C9	883
NC_018476.1	Simbu virus RdRp gene for RNA-dependent RNA polymerase	780
NC_017618.1	Streptococcus suis JS14	752
NZ_AP012496.1	Bacillus subtilis BEST7003	672
NZ_CP015375.1	Bacillus subtilis subsp, subtilis strain KCTC 3135	552
NC_019491.1	Cyprinid herpesvirus 1 strain NG-J1	461
NZ_CP008870.1	Pseudomonas aeruginosa strain W36662	423
NC_000909.1	Methanocaldococcus jannaschii DSM 2661	313
NC_009127.1	Cyprinid herpesvirus 3	306

Παρακάτω παρουσιάζονται αναλυτικά οι αφθονίες των μικροβίων που βρέθηκαν στο dataset **PRJNA285523** που λόγω της προετοιμασίας βιβλιοθήκης (16S rRNA - Seq) είναι εξ ορισμού πείραμα που στοχεύει στην αλληλούχιση των μικροβίων εντός του εντέρου των υπό εξέταση ποντικών.

Πίνακας 27 - Αποτελέσματα dataset PRJNA285523 (16S rRNA)

Dataset PRJNA285523		
Code	Strain Name	AVG Counts
NC_004917.1	Helicobacter hepaticus ATCC 51449	6422
NC_011852.1	Haemophilus parasuis SH0165	1658
NC_018011.1	Alistipes finegoldii DSM 17242	1309
NC_014393.1	Clostridium cellulovorans 743B	530
NC_014013.1	Sphingobium japonicum UT26S	197
NZ_CP012009.1	Leuconostoc mesenteroides subsp. dextranicum strain DSM 20484	189
NC_015160.1	Odoribacter splanchnicus DSM 20712	64
NC_003212.1	Listeria innocua Clip11262	40
NC_014834.1	Rhodopseudomonas palustris DX-1	31
NC_010001.1	Clostridium phytofermentans ISDg	22
NZ_CP015206.1	Pediococcus acidilactici strain ZPA017	16

Το σύνολο των reads από το sequencing των παραπάνω δειγμάτων (PRJNA285523) ήταν μόνο μερικές χιλιάδες ανά δείγμα, κι αυτό εξηγεί, εν μέρη, τον μικρό αριθμό μικροβίων και αφθονιών που εντοπίστηκαν στα δείγματα.

Η ανάλυση του συγκεκριμένου dataset έγινε με σκοπό την επικύρωση των αποτελεσμάτων των προηγούμενων dataset. Η λογική είναι πως εφόσον μιλάμε για ένα πείραμα που στόχο είχε το sequencing του μικροβιώματος του εντέρου (intestine microbiota), μπορεί στην συνέχεια να χρησιμοποιηθεί ως οδηγός για τα υπόλοιπα δείγματα και με αυτό τον τρόπο να εξηγηθεί πιο εύκολα η παρουσία κάποιων μικροβίων. Φυσικά, για τα μικρόβια που βρέθηκαν στα προηγούμενα δείγματα και όχι σε αυτό, δεν σημαίνει πως έχει γίνει απαραίτητα λάθος, απλώς η δικαιολόγηση της παρουσίας κάποιων μικροβίων ίσως είναι δυσκολότερη σε τέτοιες περιπτώσεις.

Στον πίνακα που ακολουθεί, παρουσιάζονται τα μικρόβια που εντοπίστηκαν και στα ιστοειδικά δείγματα, αλλά και στο δείγμα που είχε ως στόχο την αλληλούχιση του μικροβιώματος του εντέρου. Επίσης, στον πίνακα συμπεριλαμβάνονται και τα μικρόβια που εντοπίστηκαν στη δημοσίευση [45] η οποία είχε ως στόχο την δημιουργία ενός καταλόγου με όλα τα είδη τα οποία σχετίζονται με το μικρόβιομα του εντέρου του ποντικίου. Στη συγκεκριμένη δημοσίευση έγινε η αλληλούχιση 184 δειγμάτων περιττωμάτων από διαφορετικά ποντίκια και στην συνέχεια έγινε η παρουσίαση των βακτηρίων που εντοπίστηκαν και η σχετική αφθονία καθενός από αυτά.

Πίνακας 28 - Μικρόβια που εντοπίστηκαν σε ιστοειδικά και μικροβιακά δείγματα

Microbes identified in tissue sample datasets, intestine microbiota and [43]	
Code	Strain Name
NZ_CP014577.1	Salmonella enterica subsp. enterica serovar Typhimurium strain RM9437
NC_004917.1	Helicobacter hepaticus ATCC 51449
NC_018011.1	Alistipes finegoldii DSM 17242
NC_010001.1	Clostridium phytofermentans ISDg
NC_009668.1	Ochrobactrum anthropi ATCC 49188
NC_009720.1	Xanthobacter autotrophicus Py2
NZ_CP015725.1	Salmonella enterica strain C629
NZ_CP012141.1	Shigella flexneri 4c strain 120
NC_014976.1	Bacillus subtilis BSn5
NZ_CP011335.1	Escherichia coli O104:H4 str. C227-11
NZ_CP015837.1	Escherichia coli strain MS6198
NZ_CP015078.1	Escherichia coli strain Ecol_448
NZ_CP007392.1	Escherichia coli strain ST2747
NC_017618.1	Streptococcus suis JS14
NC_017621.1	Streptococcus suis D12
NC_004917.1	Helicobacter hepaticus ATCC 51449
NZ_CP014348.1	Escherichia coli str. K-12 substr. MG1655 strain JW5437-1
NZ_CP011113.1	Escherichia coli strain RR1
NZ_CP010439.1	Escherichia coli K-12 strain ER3440
NZ_CP015159.1	Escherichia coli strain Eco889
NZ_LN832404.1	Escherichia coli K-12 genome assembly EcoliK12AG100
NZ_CP015975.1	Bacillus subtilis subsp. subtilis strain delta6
NC_012781.1	Eubacterium rectale ATCC 33656
NC_014388.1	Butyrivibrio proteoclasticus B316
NC_009656.1	Pseudomonas aeruginosa PA7
NZ_CP013993.1	Pseudomonas aeruginosa DHS01
NC_014833.1	Ruminococcus albus 7
NZ_CP008857.1	Pseudomonas aeruginosa strain F30658
NC_015164.1	Bacteroides salanitronis DSM 18170
NZ_CP008866.1	Pseudomonas aeruginosa strain T38079
NZ_AP012496.1	Bacillus subtilis BEST7003
NC_023149.1	Pseudomonas aeruginosa SCV20265
NC_009615.1	Parabacteroides distasonis ATCC 8503
NC_018011.1	Alistipes finegoldii DSM 17242
NC_014033.1	Prevotella ruminicola 23
NC_014933.1	Bacteroides helcogenes P 36-108
NZ_CP015375.1	Bacillus subtilis subsp. subtilis strain KCTC 3135
NZ_CP008861.1	Pseudomonas aeruginosa strain H47921

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία, αναπτύχθηκε ένα υπολογιστικό εργαλείο το οποίο είναι σε θέση να εντοπίζει και να ποσοτικοποιεί μικροβιακούς οργανισμούς σε δείγματα NGS με την τεχνική του quasi-mapping.

Για την αξιολόγηση του εργαλείου, δημιουργήθηκαν συνθετικά δεδομένα αλληλούχισης επόμενης γενιάς, μεταβαλλόμενου μεγέθους και πολυπλοκότητας. Από την σύγκριση που έγινε με υπολογιστικά εργαλεία που δημιουργήθηκαν για την αντιμετώπιση του ίδιου προβλήματος, προκύπτει, πως ο αλγόριθμος που αναπτύχθηκε, έχει μεγαλύτερη ακρίβεια σε όλες τις περιπτώσεις στις οποίες δοκιμάστηκε και σε πολλές από αυτές είναι και ταχύτερος.

Συγκεκριμένα, χρησιμοποιήθηκαν πολλών ειδών μετρήσεις (metrics) ώστε τα αποτελέσματα να είναι πιο σφαιρικά, και, στις περισσότερες περιπτώσεις το εργαλείο που αναπτύχθηκε αποδείχτηκε ακριβέστερο. Σχετικά με την ταχύτητα εκτέλεσης, προκύπτει πως σε περιπτώσεις που το σύνολο των reads είναι σχετικά μικρό, το MetaPhlAn2 παράγει αποτελέσματα με μεγαλύτερη ταχύτητα, αντίθετα όταν ο αριθμός των reads είναι μεγάλος, το metaHost είναι το γρηγορότερο πρόγραμμα σε σχέση με τα προγράμματα που συμμετείχαν στις συγκρίσεις. Αξίζει να σημειωθεί πως όταν τα προς ανάλυση δείγματα προέρχονται από πειράματα Shotgun Metagenomics Sequencing, ο αριθμός των reads είναι πολύ μεγάλος και συνεπώς φαίνεται πως το metaHost σε δεδομένα που προέρχονται από πραγματικά πειράματα θα παράγει τα απαραίτητα αποτελέσματα ταχύτερα. Όσον αφορά την ακρίβεια των προγραμμάτων, το metaHost παρουσιάζει την μεγαλύτερη ακρίβεια σε όλες τις ταξινομικές βαθμίδες (γένος, είδος, υποείδος). Ακολουθεί το πρόγραμμα kallisto και τελευταία έρχονται το kraken και το MetaPhlAn2 όπου χρησιμοποιούνται και πιο συχνά σε επιστημονικές μελέτες μεταγονιδιωμικής και ανάλυσης μικροβιώματος. Ειδικά στην ταξινομική βαθμίδα των υποειδών, η ακρίβεια του metaHost είναι πρωτοφανής με το kraken να έχει εμφανή μικρότερη ακρίβεια και το MetaPhlAn2, στις περισσότερες περιπτώσεις να μην είναι καν σε θέση να υπολογίσει μικροβιακές αφθονίες στη συγκεκριμένη βαθμίδα.

Η τεχνική του quasi-mapping, αν και αρχικά αναπτύχθηκε για να λύσει το πρόβλημα της ποσοτικοποίησης των transcripts από πειράματα RNA-Seq, αποδείχθηκε πως μπορεί να εφαρμοστεί πετυχημένα και στο πρόβλημα του εντοπισμού και της ποσοτικοποίησης του μικροβιώματος σε δεδομένα μεταγονιδιωμικής και συνεπώς αποκαλύφθηκε ένα ακόμα πρόβλημα το οποίο μπορεί να επιλύσει πολύ αποδοτικά.

Ο παραπάνω αλγόριθμος, ενσωματώθηκε σε ένα pipeline το οποίο εκτός από τον εντοπισμό και την ποσοτικοποίηση των μικροβίων, υπολογίζει τις σχετικές αφθονίες στα επίπεδα του γένους, των ειδών και των υποειδών, φιλτράρει τα αποτελέσματα βάση κριτηρίων που επιλέγονται από τον χρήστη, εξάγει μερικά βασικά στατιστικά και τέλος γίνεται οπτικοποίηση αυτών σε φιλικά προς τον χρήστη γραφήματα. Όλες οι διαδικασίες είναι πλήρως αυτοματοποιημένες και τα δείγματα αναλύονται παράλληλα με σκοπό την ταχύτερη δυνατή ολοκλήρωση της εκτέλεσης του pipeline.

Τέλος, το pipeline, έχει και τη δυνατότητα ανάλυσης ιστοειδικών δειγμάτων που περιέχουν DNA/RNA του θηλαστικού οργανισμού και του μικροβιώματος που ενδεχομένως φέρει.

Μελλοντικά, μερικοί από τους στόχους που έχουν τεθεί, είναι η δημιουργία μιας ολικής βάσης με μικρόβια (reference) κάνοντας ομαδοποίηση (clustering) των μικροβιακών γονιδιωμάτων, όπου κύριος σκοπός της συγκεκριμένης βάσης θα είναι η χρησιμοποίηση της σε οποιαδήποτε μελέτη ανεξαρτήτως των δειγμάτων που αναλύονται. Στην συνέχεια, επόμενος στόχος είναι η προσθήκη του αυτόματου εντοπισμού μικροβίων με στατιστικά σημαντική διαφορά στις αφθονίες τους, μεταξύ φυσιολογικών και παθογόνων δειγμάτων και τέλος η μαζική ανάλυση πραγματικών δεδομένων με σκοπό τον εντοπισμό συσχετίσεων συγκεκριμένων μικροβίων με συγκεκριμένες ασθένειες.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός όρος
microbiome	Μικροβίωμα
genomics	Γονιδιωματική
metagenomics	Μεταγονιδιωματική
datasets	Σύνολα δεδομένων
microbiota	Μικροχλωρίδα
gram-positive	Γκραμ-θετικά
gram-negative	Γκραμ-αρνητικά
metagenome	Μεταγονιδίωμα
sequencing	Αλληλούχιση
enviromental sample	Περιβαλλοντικό δείγμα
genome	Γονιδίωμα
microbiology	Μικροβιολογία
microbial culture	Μικροβιακή καλλιέργεια
profile of diversity	Προφίλ ποικιλομορφίας
shotgun metagenomic sequencing	Μαζική αλληλούχιση μεταγονιδιώματος
identification	Εντοπισμός – Ταυτοποίηση
abundance estimation	Υπολογισμός αφθονίας
sub-species	Υπο-είδη
strains	Υπο-είδη
next generation sequencing	Αλληλούχιση επόμενης γενιάς
reference	Βάση αναφοράς
alignment	Στοίχιση
completed genome	Ολοκληρωμένο γονιδίωμα
draft genome	Γονιδίωμα υπο επεξεργασία
high similarity	Υψηλή ομοιότητα
resources	Πόροι
data storage	Αποθήκευση δεδομένων
smart incubators	Έξυπνοι επωαστήρες
conserved	Συντηρημένη
fingerprinting	Τεχνική αποτύπωσης
probe hybridization techniques	Τεχνική υβριδοποίησης ανιχνευτή
DNA microarrays	Τεχνική μικροσυστοιχιών DNA

amplicons	Αμπλικόνια
consensus sequence	Συνεχής ακολουθίες
species	Είδος
genus	Γένος
coding sequences	Ακολουθία που κωδικοποιεί πρωτεΐνη
microbial clades	Μικροβιακοί κλάδοι
taxonomic rank	Ταξινομικές βαθμίδες
relative abundance	Σχετική αφθονία
markers	Δείκτες
taxonomic labels	Ταξινομική ταμπέλα
Lower Common Ancestor	Κατώτατος κοινός πρόγονος
specific node	Συγκεκριμένος κόμβος
classification	Κατηγοριοποίηση
read	Ακολουθία που προέρχεται από δεδομένα αλληλούχησης επόμενης γενιάς
pruned subtree	Κλαδεμένο υποδέντρο
weight	Βάρος
root to leaf path	Μονοπάτι από τη ρίζα έως τα φύλλα
highest score	Ψηλότερο σκορ
Bayesian	Μπεϋζιανή
quantification	Ποσοτικοποίηση
traditional alignment	Παραδοσιακή στοίχιση ακολουθιών
pseudoalignment	Ψευδό-στοίχιση
colored de Bruijn graph	Χρωματισμένος Γράφος de Bruijn
tissue-specific samples	Ιστό-ειδικό δείγμα
sample contamination	Επιμόλυνση δείγματος
transcriptome	Μεταγράψωμα
family	Οικογένεια
rules	Κανόνες
modules	Αυτόνομη οντότητα λογισμικού
cores	Πυρήνες
mate	Ταίρι
format	Μορφή
Suffix Array	Πίνακας καταλήξεων
hash-table	Πίνακας κατατεμαχισμού

maximum mappable prefix	Μέγιστο πρόθεμα που στοιχίζεται
substrings	Υπο-αλφαριθμητικό
longest common prefix	Μέγιστο κοινό πρόθεμα
next informative position	Επόμενη θέση που δίνει πληροφορία
k-mer skipping	Παράκαμψη k-mers
plasmid	Πλασμίδιο
ground truth	Αλήθεια
prior knowledge	Πρότερη γνώση
mutations	Μεταλλάξεις
sequencing error	Λάθος αλληλούχισης
empirical error model	Μοντέλο εμπειρικού σφάλματος
technology-specific	Ειδικό στην τεχνολογία
Mouse	Ποντίκι
Mus musculus	Ποντίκι
Archaea	Αρχαία
metrics	Μετρικές

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

RAM	Random access memory
NGS	Next generation sequencing
LCA	Lowest common ancestor
LCP	Longest common prefix
NIP	Next informative position
SA	Suffix array
MMP	Maximum mappable prefix
SAM	Sequence alignment/map format
NCBI	National center for biotechnology information

ΑΝΑΦΟΡΕΣ

- [1] M. Madigan, J. Martinko and J. Parker, Brock biology of microorganisms, 1st ed. Upper Saddle River, NJ: Prentice Hall/Pearson Education, 2003.
- [2] "First-Ever Scientific Estimate of Total Bacteria on Earth Shows Far Greater Numbers than Ever Known Before", ScienceDaily, 2017. [Online]. Available: <https://www.sciencedaily.com/releases/1998/08/980825080732.htm>. [Accessed: 16- Jan- 2017].
- [3] K. Dose, A. Bieger-Dose, R. Dillmann, M. Gill, O. Kerz, A. Klein, H. Meinert, T. Nawroth, S. Risi and C. Stridde, "ERA-experiment "space biochemistry"", Advances in Space Research, vol. 16, no. 8, pp. 119-129, 1995.
- [4] ' . Words, "'Ome Sweet 'Omics-- A Genealogical Treasury of Words | The Scientist Magazine®", The Scientist, 2017. [Online]. Available: <http://www.the-scientist.com/?articles.view/articleNo/13313/title/-Ome-Sweet--Omics---A-Genealogical-Treasury-of-Words/>. [Accessed: 16- Jan- 2017].
- [5] J. Faith, P. Ahern, V. Ridaura, J. Cheng and J. Gordon, "Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice", Science Translational Medicine, vol. 6, no. 220, pp. 220ra11-220ra11, 2014.
- [6] J. Rosner, "Ten Times More Microbial Cells than Body Cells in Humans?", Microbe Magazine, vol. 9, no. 2, pp. 47-47, 2014.
- [7] S. Luria, General virology, 1st ed. New York: Wiley, 1978.
- [8] J. Handelsman, M. Rondon, S. Brady, J. Clardy and R. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products", Chemistry & Biology, vol. 5, no. 10, pp. R245-R249, 1998.
- [9] K. Schulze-Schweifing, A. Banerjee and W. Wade, "Comparison of bacterial culture and 16S rRNA community profiling by clonal analysis and pyrosequencing for the characterization of the dentine caries-associated microbiome", Frontiers in Cellular and Infection Microbiology, vol. 4, 2014.
- [10] J. Lagier, F. Armougom, M. Million, P. Hugon, I. Pagnier, C. Robert, F. Bittar, G. Fournous, G. Gimenez, M. Maraninchi, J. Trape, E. Koonin, B. La Scola and D. Raoult, "Microbial culturomics: paradigm shift in the human gut microbiome study", Clinical Microbiology and Infection, vol. 18, no. 12, pp. 1185-1193, 2012.
- [11] J. Rajendhran and P. Gunasekaran, "Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond", Microbiological Research, vol. 166, no. 2, pp. 99-110, 2011.
- [12] J. Janda and S. Abbott, "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls", Journal of Clinical Microbiology, vol. 45, no. 9, pp. 2761-2764, 2007.
- [13] P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. Sogin, W. Jones, B. Roe, J. Affourtit, M. Egholm, B. Henrissat, A. Heath, R. Knight and J. Gordon, "A core gut microbiome in obese and lean twins", Nature, vol. 457, no. 7228, pp. 480-484, 2008.
- [14] J. Venter, "Environmental Genome Shotgun Sequencing of the Sargasso Sea", Science, vol. 304, no. 5667, pp. 66-74, 2004.
- [15] T. Nawy, "Microbiome: Expanding the Gut Gene Catalog", PSI Structural Genomics KnowledgeBase, 2014.
- [16] "NIH Human Microbiome Project", Microbe Magazine, vol. 4, no. 9, pp. 393-393, 2009.
- [17] S. Hiraoka, A. Machiyama, M. Ijichi, K. Inoue, K. Oshima, M. Hattori, S. Yoshizawa, K. Kogure and W. Iwasaki, "Genomic and metagenomic analysis of microbes in a soil environment affected by the 2011 Great East Japan Earthquake tsunami", BMC Genomics, vol. 17, no. 1, 2016.
- [18] N. Fierer, J. Leff, B. Adams, U. Nielsen, S. Bates, C. Lauber, S. Owens, J. Gilbert, D. Wall and J. Caporaso, "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes", Proceedings of the National Academy of Sciences, vol. 109, no. 52, pp. 21390-21395, 2012.
- [19] S. Colston, M. Fullmer, L. Beka, B. Lamy, J. Gogarten and J. Graf, "Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using Aeromonas as a Test Case", mBio, vol. 5, no. 6, pp. e02136-14, 2014.
- [20] T. Větrovský and P. Baldrian, "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses", PLoS ONE, vol. 8, no. 2, p. e57923, 2013.

- [21] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes", *Nature Methods*, vol. 9, no. 8, pp. 811-814, 2012.
- [22] D. Truong, E. Franzosa, T. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower and N. Segata, "Erratum: MetaPhlAn2 for enhanced metagenomic taxonomic profiling", *Nature Methods*, vol. 13, no. 1, pp. 101-101, 2015.
- [23] D. Wood and S. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments", *Genome Biology*, vol. 15, no. 3, p. R46, 2014.
- [24] R. Ounit, S. Wanamaker, T. Close and S. Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers", *BMC Genomics*, vol. 16, no. 1, 2015.
- [25] J. Lu, F. Breitwieser, P. Thielen and S. Salzberg, "Bracken: estimating species abundance in metagenomics data", *PeerJ Computer Science*, vol. 3, p. e104, 2017.
- [26] N. Bray, H. Pimentel, P. Melsted and L. Pachter, "Erratum: Near-optimal probabilistic RNA-seq quantification", *Nature Biotechnology*, vol. 34, no. 8, pp. 888-888, 2016.
- [27] L. SCHAEFFER, H. PIMENTEL, N. BRAY, P. MELSTED and L. PACHTER, "PSEUDOALIGNMENT FOR METAGENOMIC READ ASSIGNMENT", 2016.
- [28] "Corrections to Papers: Maximum Likelihood Estimation from Incomplete Data", *Biometrics*, vol. 14, no. 4, p. 562, 1958.
- [29] A. Monnappa, W. Bari, S. Choi and R. Mitchell, "Investigating the Responses of Human Epithelial Cells to Predatory Bacteria", *Scientific Reports*, vol. 6, p. 33485, 2016.
- [30] Y. Han, W. Shi, G. Huang, S. Kinder Haake, N. Park, H. Kuramitsu and R. Genco, "Interactions between Periodontal Bacteria and Human Oral Epithelial Cells: *Fusobacterium nucleatum* Adheres to and Invades Epithelial Cells", *Infection and Immunity*, vol. 68, no. 6, pp. 3140-3146, 2000.
- [31] R. Reichert and G. Stern, "Quantitative Adherence of Bacteria to Human Corneal Epithelial Cells", *Archives of Ophthalmology*, vol. 102, no. 9, pp. 1394-1395, 1984.
- [32] G. Melmed, L. Thomas, N. Lee, S. Tesfay, K. Lukasek, K. Michelsen, Y. Zhou, B. Hu, M. Ardit and M. Abreu, "Human Intestinal Epithelial Cells Are Broadly Unresponsive to Toll-Like Receptor 2-Dependent Bacterial Ligands: Implications for Host-Microbial Interactions in the Gut", *The Journal of Immunology*, vol. 170, no. 3, pp. 1406-1415, 2003.
- [33] M. Strong, G. Xu, L. Morici, S. Splinter Bon-Durant, M. Baddoo, Z. Lin, C. Fewell, C. Taylor and E. Flemington, "Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples", *PLoS Pathogens*, vol. 10, no. 11, p. e1004437, 2014.
- [34] M. Laurence, C. Hatzis and D. Brash, "Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes", *PLoS ONE*, vol. 9, no. 5, p. e97876, 2014.
- [35] J. Koster and S. Rahmann, "Snakemake--a scalable bioinformatics workflow engine", *Bioinformatics*, vol. 28, no. 19, pp. 2520-2522, 2012.
- [36] M. Pertea, D. Kim, G. Pertea, J. Leek and S. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown", *Nature Protocols*, vol. 11, no. 9, pp. 1650-1667, 2016.
- [37] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin, "The Sequence Alignment/Map format and SAMtools", *Bioinformatics*, vol. 25, no. 16, pp. 2078-2079, 2009.
- [38] F. Sanger, G. Air, B. Barrell, N. Brown, A. Coulson, J. Fiddes, C. Hutchison, P. Slocombe and M. Smith, "Nucleotide sequence of bacteriophage ϕ X174 DNA", *Nature*, vol. 265, no. 5596, pp. 687-695, 1977.
- [39] A. Srivastava, H. Sarkar, N. Gupta and R. Patro, "RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes", *Bioinformatics*, vol. 32, no. 12, pp. i192-i200, 2016.
- [40] R. Patro, G. Duggal and C. Kingsford, "Accurate, fast, and model-aware transcript expression quantification with Salmon", 2016.
- [41] R. Patro, S. Mount and C. Kingsford, "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms", *Nature Biotechnology*, vol. 32, no. 5, pp. 462-464, 2014.
- [42] J. Hunter, "Matplotlib: A 2D Graphics Environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007.
- [43] L. Manley, D. Ma and S. Levine, "Monitoring Error Rates in Illumina Sequencing", *Journal of Biomolecular Techniques: JBT*, pp. jbt.16-2704-002, 2016.

- [44] W. Huang, L. Li, J. Myers and G. Marth, "ART: a next-generation sequencing read simulator", *Bioinformatics*, vol. 28, no. 4, pp. 593-594, 2011.
- [45] L. Xiao, Q. Feng, S. Liang, S. Sonne, Z. Xia, X. Qiu, X. Li, H. Long, J. Zhang, D. Zhang, C. Liu, Z. Fang, J. Chou, J. Glanville, Q. Hao, D. Kotowska, C. Colding, T. Licht, D. Wu, J. Yu, J. Sung, Q. Liang, J. Li, H. Jia, Z. Lan, V. Tremaroli, P. Dworzynski, H. Nielsen, F. Bäckhed, J. Doré, E. Le Chatelier, S. Ehrlich, J. Lin, M. Arumugam, J. Wang, L. Madsen and K. Kristiansen, "A catalog of the mouse gut metagenome", *Nature Biotechnology*, vol. 33, no. 10, pp. 1103-1108, 2015.
- [46] T. Nguyen, S. Vieira-Silva, A. Liston and J. Raes, "How informative is the mouse for human gut microbiota research?", *Disease Models & Mechanisms*, vol. 8, no. 1, pp. 1-16, 2015.
- [47] T. Clavel, I. Lagkouvardos, M. Blaut and B. Stecher, "The mouse gut microbiome revisited: From complex diversity to model ecosystems", *International Journal of Medical Microbiology*, vol. 306, no. 5, pp. 316-327, 2016.
- [48] D. Laukens, B. Brinkman, J. Raes, M. De Vos and P. Vandenabeele, "Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design", *FEMS Microbiology Reviews*, vol. 40, no. 1, pp. 117-132, 2015.