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*:** **Αρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.

ΑΘΗΝΑ

ΙΑΝΟΥΑΡΙΟΣ 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

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

Γιώργος Ν. Σκούφος

Α.Μ.: ΠΙΒ0145

ΕΠΙΒΛΕΠΟΥΣΑ: **Αρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Αρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.
Ιωάννης Βλάχος, Ερευνητής Ιατρικής Σχολής του Πανεπιστημίου Χάρβαρντ
Μαρία Παρασκευοπούλου, Ερευνήτρια/Επιστημονικός Συνεργάτης

Ιανουάριος 2017

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη των τεχνολογιών αλληλούχισης επόμενης γενιάς (Next Generation Sequencing - NGS) έχουν μετατρέψει την ικανότητα μας για διερεύνηση της σύνθεση 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 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 ποσοτικοποιήσεις 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 πλήρης 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 human skin, gut and oral. Sequenced metagenomic samples usually comprise reads from a large number of different bacterial and viral communities and hence tend to result in huge 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 microbiota.

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