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 των μικρών μορίων RNA από δεδομένα αλληλούχησης επόμενης γενιάς με τη χρήση του spipeRNA

Joanna A. Handzlik

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

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση των μικρών RNA Ανάλυση των μικρών μορίων RNA από δεδομένα αλληλούχησης επόμενης γενιάς με τη χρήση του spipeRNA

Joanna A. Handzlik
A.M.: ΠΙΒ0140

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

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Άρτεμις Χατζηγεωργίου**, Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Υ/Η, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας
Γεώργιος Μ. Σπύρου, Επικεφαλής της ομάδας Βιοπληροφορικής, Κυπριακό Ινστιτούτου Νευρολογίας και Γενετικής
Ιωάννης Βλάχος, Ερευνητής Ιατρικής Σχολής Πανεπιστημίου Χάρβαρντ

Δεκέμβριος 2016

ΠΕΡΙΛΗΨΗ

Μία από τις σημαντικότερες τεχνολογικές εξελίξεις στο χώρο της βιοτεχνολογίας τα τελευταία χρόνια, αφορά την τεχνολογία της αλληλούχησης επόμενης γενιάς (Next Generation Sequencing ή NGS). Ενώ η αποκρυπτογράφηση του πρώτου ανθρωπίνου γονιδιώματος (3 δις βάσεις ανά απλοειδές γονιδίωμα) χρειάστηκε περίπου 15 χρόνια με υλικό κόστος και μόνο, ανερχόμενο στα 10 δις δολάρια Αμερικής, σήμερα η αλληλούχηση ολόκληρου του ανθρωπίνου γονιδιώματος μπορεί να παραχθεί από μία μονάχα συσκευή μέσα σε λίγες μέρες, με κόστος που δεν ξεπερνά τα 1.000 δολάρια Αμερικής. Η τεχνολογία αυτή άνοιξε το δρόμο για πολλές συναρπαστικές εφαρμογές, όπως είναι η *de novo* αλληλούχηση, η ανάλυση του μεταγραφώματος (RNA-Seq) και του μεθυλώματος (methyl-seq), ο προσδιορισμός των θέσεων πρόσδεσης των μεταγραφικών παραγόντων (ChIP-seq), η ανίχνευση των γονιδιακών μεταλλάξεων υπεύθυνων για ασθένειες και πολλές άλλες ακόμη εφαρμογές.

Ο σκοπός της παρούσας διπλωματικής ήταν ο σχεδιασμός και η υλοποίηση ενός υπολογιστικού εργαλείου αφιερωμένου στην ανάλυση των small RNA-Seq δεδομένων, τα οποία αποτελούν ένα κομμάτι της γενικής ανάλυσης του μεταγραφώματος (RNA-Seq). Η ανάλυση αυτή αποσκοπεί στην ποσοτικοποίηση των εκφραζόμενων μικρών μορίων RNA και την εύρεση καινούργιων, μη σχολιασμένων περιοχών έκφρασης, σε βιολογικά δείγματα ποικίλλης προέλευσης.

Ο αλγόριθμος που σχεδιάστηκε, ονομάστηκε “spipeRNA» και απαντά σε πολλές ανοιχτές προκλήσεις: ποσοτικοποιεί όλα τα μικρά RNAs και όχι μόνο τα miRNAs, επιλύει το πρόβλημα των εγγραφών με πολλαπλές θέσεις ευθυγράμμισης πάνω στο γονιδίωμα και χειρίζεται κατάλληλα τις εγγραφές χωρίς υπάρχοντα σχολιασμό.

Στην εργασία παρουσιάζονται και τα αποτελέσματα εκτέλεσης του εργαλείου για την ανάλυση προσομοιωμένων δεδομένων και 8 small RNA-Seq συνόλων δεδομένων, τα οποία περιλαμβάνουν καρκινικά και υγιή δείγματα πνεύμονα και παγκρέατος. Το spipeRNA συγκρίθηκε με ένα δημοφιλές εργαλείο ανάλυσης των miRNAs επιδεικνύοντας υψηλότερη ακρίβεια σε προσομοιωμένα και πραγματικά δεδομένα.

Το εργαλείο spipeRNA, βασίζεται σε μία αξιόπιστη, ευέλικτη και πλήρως αυτοματοποιημένη ροή εργασιών, χρήσιμη για τη γρήγορη και υψηλής απόδοσης ανάλυση των small RNA-Seq δεδομένων από αλληλουχητές επόμενης γενιάς.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μικρά μη-κωδικά RNAs, αλληλούχηση επόμενης γενιάς, NGS, ευθυγράμμιση πάνω στο γονιδίωμα, σχολιασμός γονιδιωματικών περιοχών, microRNA, snoRNA, snRNA, tRNA, rRNA, siRNA

ABSTRACT

Some of the most important technological developments in biotechnology in recent years, are summarized under the term “Next Generation Sequencing (NGS)”. While the sequencing of the first human genome (3 gigabases per haploid genome) took about 15 years and roughly 100 million of US dollars in material costs only, today the raw sequencing data for a complete human genome (100 gigabases at 30x coverage) can be produced by a single machine within a few days and for just 1.000 US dollars. This technological quantum leap has paved the way for numerous exciting applications such as de novo sequencing, transcriptome (RNA-Seq) and methylome (methyl-Seq) analysis, the determination of transcription factor binding sites (ChIP-Seq), the detection of disease-causing mutations, and many others.

The purpose of this study was the design and implementation of the computational tool, dedicated to the analysis of small RNA-Seq data, which form a part of the overall analysis of trascriptome (RNA-Seq). This analysis aims to quantify the expressed small RNA molecules and to detect new non-annotated expression regions in various biological samples.

The implemented algorithm was called "spipeRNA» and tries to overcome many open challenges: it quantifies all types of small RNAs, not only the miRNAs, solves the problem of multi-mapped reads and appropriately handles the reads without existing annotation.

This study presents the results obtained by applying this tool to analyze simulated data and 8 small RNA-Seq datasets, which include tumor/healthy lung and pancreas samples. The comparison between the spipeRNA and very popular tool for miRNAs analysis, showed, that in some cases, the spipeRNA may produce more precise and accurate output.

The spipeRNA is an integrated data analysis pipeline, based on a reliable, flexible and fully automated workflow, useful for fast and efficient analysis of small RNA-Seq data produced by next-generation sequencers.

.

SUBJECT AREA: Bioinformatics

KEYWORDS: small non-coding RNAs, Next Generation Sequencing, NGS, reads alignment, annotation of genomic regions, microRNA, snoRNA, snRNA, tRNA, rRNA, siRNA