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

Διαχείριση κι ανάλυση δεδομένων, υψηλής απόδοσης, με στόχο την εύρεση και κατανόηση γενετικών μοριακών αλλαγών που σχετίζονται με τον καρκίνο του μαστού

Δήμητρα Ε. Καραγκούνη

Επιβλέπουσα: Άρτεμις Χατζηγεωργίου, Καθηγήτρια

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2014

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διαχείριση κι ανάλυση δεδομένων, υψηλής απόδοσης, με στόχο την εύρεση και κατανόηση γενετικών μοριακών αλλαγών που σχετίζονται με τον καρκίνο του μαστού

Δήμητρα Ε. Καραγκούνη
ΠΙΒ: 0104

Επιβλέπουσα: Άρτεμις Χατζηγεωργίου, Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: Άρτεμις Χατζηγεωργίου, Καθηγήτρια
Γεώργιος Σπύρου, Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄)
Χρυσίνα Ευαγγελία, Επιστημονικό Προσωπικό του ΕΙΕ

ΙΟΥΝΙΟΣ 2014

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική βασίζεται σε καρκινικά δεδομένα μελέτης που έχει πραγματοποιηθεί από το δίκτυο του Cancer Genome Atlas (TCGA: Comprehensive molecular portraits of human breast tumours,Nature 2012). Τα δεδομένα αυτά επικεντρώνονται στην ανάλυση του γονιδιώματος με υψηλής απόδοσης τεχνικές κι έχουν ως στόχο την καλύτερη κατανόηση της γενετικής βάσης του καρκίνου. Η δική μας μελέτη εστιάζει στη διαχείριση και ανάλυση τέτοιων δεδομένων, υψηλής διεκπεραιωτικής ικανότητας, στους βασικούς τύπους καρκίνου του μαστού.

Το πρώτο σκέλος της διπλωματικής αναφέρεται στη δημιουργία ενός αυτοματοποιημένου προγράμματος, το οποίο μπορεί να χρησιμοποιηθεί σε δεδομένα έκφρασης, για την εύρεση ιδανικών γονιδίων που προσφέρουν την καλύτερη διαχωριστική ικανότητα σε διαφορετικούς τύπους μίας οποιασδήποτε ασθένειας. Συγκεκριμένα χρησιμοποιούνται μέθοδοι εύρεσης των μεταγράφων-γονιδίων με την μεγαλύτερη μεταβλητότητα έκφρασης στο σύνολο των δειγμάτων που οδηγούν στην καλύτερη κατηγοριοποίηση τους. Με βάση την παραπάνω ομαδοποίηση βρίσκονται οι πιο διαφοροποιημένοι στην έκφρασή τους ρυθμιστές και κατ' επέκταση πιο αντιπροσωπευτικοί ανά ομάδα. Στην παρούσα μελέτη καταλήξαμε σε ομάδες γονιδίων που παρουσιάζουν την πιο μεταβλητή έκφραση στους βασικούς τύπους καρκίνου του μαστού.

Το δεύτερο σκέλος της διπλωματικής εργασίας ασχολείται με την ανάλυση των μεταλλαγών των γονιδίων, ενός ή περισσότερων νουκλεοτιδίων σε περιοχές πρόσδεσης των μικρών μη κωδικών RNAs, miRNAs, μετά-μεταφραστικών ρυθμιστών που παίζουν καθοριστικό ρόλο στη ρύθμιση της έκφρασης των περισσότερων γονιδίων. Από την παραπάνω ανάλυση παρατηρείται ότι υπάρχουν miRNAs που προσδένονται στο αγγελιαφόρο RNA γονιδίων που έχουν μία ή περισσότερες μεταλλαγές βάσεων στις περιοχές αυτές σ' ένα αρκετά μεγάλο αριθμό δειγμάτων. Η μεταλλαγή των βάσεων μπορεί να έχει ως αποτέλεσμα τη μη πρόσδεση των συγκεκριμένων miRNAs ή ακόμα και την πρόσδεση άλλων μικρών μη κωδικών RNAs στις περιοχές αυτές, γεγονός που μπορεί να συντελέσει στη διαφοροποίηση της έκφρασης αυτών των γονιδίων στους συγκεκριμένους τύπους καρκίνου του μαστού σε σχέση με τα φυσιολογικά δείγματα.

Η μελέτη όλου αυτού του ρυθμιστικού κυκλώματος γονιδίων, μεταλλαγών, μετα-μεγραφικών ρυθμιστών μπορεί να συνδέεται με αλλαγές σε μοριακά μονοπάτια που εμπλέκονται με τον καρκίνο και με την όποια υπό μελέτη παθολογική κατάσταση.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: καρκίνος του μαστού, δεδομένα υψηλής απόδοσης, χαρακτηριστικά γονίδια, microRNA, πρόβλεψη στόχων, μεταλλαγές βάσεων.

ABSTRACT

This thesis is based on tumor data carried out by the Cancer Genome Atlas network (TCGA: Comprehensive molecular portraits of human breast tumours, Nature 2012). These data focused on genome analysis by high performance techniques with the aim to better understand the genetic basis of cancer. Our study focuses on the management and analysis of such high throughput data in the main types of breast cancer.

The first part of this thesis refers to the creation of an automated program, which can be used in expression data for finding significant genes which offer the ideal separation of different types of any disease. Specifically, methods for finding gene-transcripts with greater variability of expression in all samples were used and lead to better categorization. Based on this classification the most diverse in their expression regulators were found, and they were characterized as the most representative per group. In the present study we came in groups of genes that exhibit the most variable expression in the major types of breast cancer.

The second part of the thesis deals with the analysis of genes mutations, one or more nucleotides in binding regions of small non-coding RNAs, miRNAs, which play a key role in regulating the expression of most genes. From the above analysis miRNAs, which bind to mutated messenger RNA genes, were found. The mutations referred to MRE regions in quite a large number of samples. SNPs can lead to non-specific binding of miRNAs or even the binding of other small non-coding RNAs in these regions, may contribute to the modulation of expression of these genes to specific types of breast cancer compared with normal samples.

The study of the entire regulatory genes network, mutations and miRNAs can be associated with changes in molecular pathways involved in cancer, and any other under investigation pathological condition.

SUBJECT AREA: Computational Biology

KEYWORDS: breast cancer, high- throughput data, significant genes, miRNA, target prediction, SNPs