ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΕΣ ΜΕΘΟΔΟΙ ΕΥΡΕΣΗΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗΣ ΥΠΟΨΗΦΙΩΝ ΓΟΝΙΔΙΩΝ ΓΙΑ ΜΕΛΕΤΕΣ ΓΕΝΕΤΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ

Λούης Παπαγεωργίου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Μπάγκος Παντελής
Επίκουρος Καθηγητής

Λαμία, 2010

ΠΕΡΙΕΧΟΜΕΝΑ

Αφιερωμένο στους αγαπημένους μου γονείς, στα αγαπητά μου αδέρφια και στην μακαριστή θεία μου Δέσποινα που με ώθησε να ενταχθώ στην πανεπιστημιακή κοινότητα και να αρχίσω την έρευνα μου.

Πρόλογος

Η παρούσα διπλωματική εργασία με θέμα «Αυτοματοποιημένες μέθοδοι εύρεσης και ταξινόμησης υποψηφίων γονιδίων για μελέτες γενετικής συσχέτισης», πραγματοποιήθηκε εξ ολοκλήρου στο Τμήμα Πληροφορικής με εφαρμογές στην Βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδας, κατά τη διάρκεια του ακαδημαϊκού έτους, 2009-2010, υπό την επίβλεψη του Επίκουρου Καθηγητή, κ. Παντελή Μπάγκου.

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου για την εμπιστοσύνη που μου έδειξε κατά την ανάθεση της διπλωματικής εργασίας, αλλά και για το μεγάλο ζήλο και όρεξη που έδειξε κατά όλη την χρονική διάρκεια διεκπεραίωσης της πτυχιακής μου, παρέχοντας μου πολύτιμη βοήθεια, χρήσιμες πληροφορίες και συμβουλές .

Θερμά ευχαριστώ στον καθηγητή κ. Δελήμπαση, για τις γόνιμες συζητήσεις μας πάνω σε θέματα προγραμματιστικού μέρους της πτυχιακής μου, την συμπαράσταση του και το ενδιαφέρον του για τη σωστή λειτουργία του κώδικα μου.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένειά μου, αλλά και στους φίλους μου που βρίσκονταν πάντα δίπλα μου με αγάπη και ένα πλατύ χαμόγελο.

Περίληψη

Ένα από τα σημαντικότερα επιτεύγματα της Βιοπληροφορικής είναι η δημιουργία των έξυπνων ιεραρχικών μηχανών αναζήτησης για υποψήφια γονίδια. Εξαιτίας του μεγάλου όγκου σε βιολογικά δεδομένα που απαρτίζουν σήμερα τις βιολογικές βάσεις δεδομένων, αν δεν υπήρχαν οι έξυπνες μηχανές ιεραρχικής αναζήτησης, τότε όλα τα δεδομένα που περιέχουν θα ήταν άχρηστα, αφού ο κάθε ερευνητής θα έπρεπε να ψάχνει στα δεδομένα των βάσεων για μία ολόκληρη ζωή, ούτως ώστε να ανακαλύψει αυτό που επιθυμεί. Πέρα από τα ταξινομημένα αποτελέσματα που μας προσφέρουν οι ιεραρχικοί αλγόριθμοι αναζήτησης, είναι ικανοί να μας παρέχουν μία γκάμα από διαφορετικές τεχνικές και μεθόδους συσχετίσεις των αποτελεσμάτων τους. Αυτό είναι πολύ σημαντικό, γιατί ο ερευνητής μπορεί να έχει μία πιο σφαιρική εικόνα γύρω από το αντικείμενο που αναζητά.

Στην παρούσα εργασία στο πρώτο μέρος ασχολούμαστε με τους ιεραρχικούς αλγόριθμους εύρεσης και ταξινόμησης υποψηφίων γονιδίων που λειτουργούν απλά με την εισαγωγή των λέξεων-κλειδιών που σχετίζονται με ασθένειες. Στο δεύτερο μέρος δημιουργούμε τεχνικές για τον συνδυασμό των αποτελεσμάτων τους. Έπειτα εισάγουμε τα συνδυασμένα αποτελέσματα τους σε δύο μεθόδους στατιστικής ανάλυσης, οι οποίες μας παρουσιάζουν τα υψηλά υποψήφια γονίδια που σχετίζονται με τις ασθένειες που αναζητούμε. Μέσα από την πορεία της όλης διαδικασίας, παράγουμε δεδομένα κάνοντας χρήση πολλών τεχνικών/μεθόδων, αλλά και διαφορετικού τύπου βιολογικών παραμέτρων, έτσι τα αποτελέσματα μας να είναι πιο αξιόπιστα από εκείνα ενός ιεραρχικού αλγόριθμου αναζήτησης. Η εργασία εκπονήθηκε για τις ασθένειες Καρκίνο του μαστού, Διαβήτη τύπου Ι, Διαβήτη τύπου ΙΙ, Υπέρταση, Σκλήρυνση κατά πλάκα και Παχυσαρκία.

Λέξεις κλειδιά

Ιεραρχικοί αλγόριθμοι αναζήτησης, μεθόδοι στατιστικής ανάλυσης, υποψήφια γονίδια, λέξης κλειδιά, ιατροβιολογικές βάσεις δεδομένων, βιολογικά δεδομένα, βιολογικές παράμετροι, Καρκίνος του μαστού, Διαβήτης τύπου Ι, Διαβήτης τύπου ΙΙ, Υπέρταση, Σκλήρυνση κατά πλάκα, Παχυσαρκία

Abstract

One of the greatest achievements of bioinformatics is the creation of a hierarchy of intelligent engines for the search of candidate genes. Through the gap of the very large scale in biomedical data that make up today biomedical databases, if the absence of hierarchical intelligent search engines would exist, then all the information they contain would be useless, since each researcher will have to look to data bases for a lifetime in order to discover what the researcher seeks. Besides the hierarchical search we are offered by a hierarchical search algorithm, this is also able to offer us a range of different techniques and methods of correlation effects. This is very important because the researcher can have a more comprehensive picture around the object that is searching for.

In the present study, the first part is dealt with the hierarchical algorithm finding and sorting candidate genes that simply operate by inserting the keyword-related diseases. In the second part, we created techniques for combining the results and subsequent introduction of combined results of two methods of statistical analysis which they have high candidate genes associated with the diseases we are looking for. Through the course of the process, we generate data using several techniques / methods and different types of biological parameters so our results to be far more reliable than the results of a hierarchical search algorithm. The work was performed for diseases of breast cancer, type I diabetes, type II diabetes, hypertension, multiple sclerosis, and obesity.

Keywords

Hierarchical search algorithm, methods of statistical analysis, candidate genes, keywords, biomedical databases, biomedical data, biological parameters, breast Cancer, Type I diabetes, Type II Diabetes, hypertension, multiple sclerosis, obesity

ΚΕΦΑΛΑΙΟ 1: ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

1. Εισαγωγή

Με βάση τα δεδομένα της σημερινής εποχής, όπου έχουμε μία ραγδαία ανάπτυξη της τεχνολογίας και γενικότερα όλων των θετικών επιστημών, δημιουργήθηκαν καινούργιες επιστήμες όπως η Βιοπληροφορική. Σιγά σιγά μέσα από την πορεία ανάπτυξης της Βιοπληροφορικής, που ολοένα και κατακτά περισσότερο έδαφος στον τομέα των θετικών επιστημών, αλλά και των επιστημών υγείας, βλέπουμε τους επιστήμονες / ερευνητές να αποφεύγουν τις κλασσικές χρονοβόρες αλλά και δαπανηρές πειραματικές μεθόδους, ανίχνευσης, συσχέτισης, πρόγνωσης τρισδιάστατης δομής, εύρεσης γενικών χαρακτηριστικών σε μοριακό και λειτουργικό επίπεδο των πρωτεϊνών, γονιδίων και DNA ακολουθιών.

Οι υπάρχουσες βάσεις δεδομένων που έχουν δημιουργηθεί συνεχώς μεγαλώνουν ως προς τον αριθμό των δεδομένων που απαριθμούν αλλά και ως προς την αξιοπιστία των δεδομένων που μας προσφέρουν. Μέσα από αυτές ο κάθε ερευνητής μπορεί να ανατρέξει και να ανακτήσει τα δεδομένα στο πεδίο που τον ενδιαφέρει, αλλά και με τη χρήση των διαφόρων έξυπνων υπολογιστικών μεθόδων που υπάρχουν να τα αναλύσει και να εξάγει εύκολα και γρήγορα τα δικά του αποτελέσματα.

Σκοπός της παρούσας εργασίας είναι να συνδέσει όλες τις μεθόδους / αλγόριθμους εύρεσης και ιεράρχησης υποψηφίων γονιδίων, που σχετίζονται με ασθένειες και υπάρχουν σήμερα διαθέσιμες στο διαδίκτυο, για να παράγουμε και να δημιουργήσουμε τα δικά μας δεδομένα μέσα από τον συνδυασμό των δεδομένων που μας παρέχουν για τις ασθένειες που μελετούσαμε. Οι ασθένειες που μελετήθηκαν στην εργασία μας είναι ο καρκίνος του μαστού, διαβήτης τύπου I-II, παχυσαρκία, σκλήρυνση κατά πλάκα και υπέρταση.

Τα δικά μας δεδομένα, με τα υποψήφια γονίδια που παράχθηκαν μέσα από το συνδυασμό των δεδομένων από τους διάφορους αλγόριθμους αναζήτησης και ιεράρχησης υποψηφίων γονιδίων που σχετίζονται με ασθένειες, εισάγονται σε δύο άλλες υπολογιστικές μεθόδους στατιστικής ανάλυσης, το RankProd[1] και Metradisc[2], ούτως ώστε μέσα από τις στατιστικές αναλύσεις που μας παρέχουν οι δύο αυτές μέθοδοι να μπορέσουμε να παρουσιάσουμε με μεγαλύτερη αξιοπιστία τα πιο υψηλά πιθανά γονίδια που σχετίζονται με την κάθε μία ασθένεια που αναζητούμε ξεχωριστά.

Έγινε προσπάθεια συνδυασμού υψηλών υποψήφιων γονιδίων που σχετίζονται με δύο ασθένειες και εξαγωγή κοινών γονιδίων που παρουσιάζονται σε δυο ασθένειες που έχουν κάποια κοινά χαρακτηριστικά γνωρίσματα.

Είναι αξιοσημείωτο να σημειωθεί ότι η όλη έρευνα και εργασία μας είναι η βάση για την δημιουργία ενός πολυεργαλείου αναζήτησης υποψηφίων γονιδίων που σχετίζονται με ασθένειες. Συνδυάζει όλους τους ελεύθερα διαθέσιμους υπολογιστικούς τρόπους αναζήτησης που υπάρχουν μέχρι σήμερα στο διαδίκτυο. Ο καθένας παράγει τα δικά του αποτελέσματα με διαφορετικές τεχνικές, αλλά και χρήση δεδομένων από διαφορετικές βάσεων δεδομένων, με αποτέλεσμα τα δεδομένα που παράγονται να προέρχονται από μία μεγάλη γκάμα βάσεων δεδομένων αλλά και διαφορετικών τεχνικών βαθμολόγησης των υποψηφίων γονιδίων.

1.1 Μέθοδοι / Αλγόριθμοι

1.1.1 Αλγόριθμος PhenoPred.

Ένα από τα πιο σημαντικά καθήκοντα της σύγχρονης Βιοπληροφορικής είναι η ανάπτυξη υπολογιστικών εργαλείων που να μπορούν να χρησιμοποιηθούν για την εύρεση και κατανόηση των ανθρώπινων νόσων. Μέχρι σήμερα μία μεγάλη γκάμα από αλγορίθμους εύρεσης και ιεράρχησης υποψήφιων γονιδίων που σχετίζονται με ασθένειες έχουν διερευνηθεί.[3-12] Ο αλγόριθμος Phenopred[13] κατασκευάστηκε για την ανίχνευση και ιεράρχηση των ανθρώπινων γονιδίων που σχετίζονται με ασθένειες με βάση τα ανθρώπινα δίκτυα αλληλεπιδράσεων πρωτεϊνικών, τα γνωστά συσχετισμένα γονίδια με ασθένειες, τις ακολουθίες πρωτεϊνών και τις λειτουργικές τους πληροφορίες σε μοριακό επίπεδο[14, 15]. Η μέθοδός , PhenoPred, εποπτεύει:

- Αρχικά για κάθε γονίδιο χαρτογραφούνται οι πρωτεΐνες που σχετίζονται με την νόσο που αναζητούμε. Η λειτουργικότητα τους βασίζεται στην απόσταση όλων των σχολιασμένων πρωτεϊνών με βάση το δίκτυο αλληλεπίδρασης, της κωδικοποιημένης ακολουθίας, των φυσικοχημικών ιδιοτήτων τους και των προβλεπόμενων δομικών ιδιοτήτων τους όπως η δευτεροταγής δομή και ευελιξία[16-18].

- Στη συνέχεια, με την βοήθεια των διανυσματικών μηχανών στήριξης [19, 20] αναζητούνται τα υποψήφια γονίδια για μια σειρά όρων - λέξεων κλειδιών που αντιπροσωπεύουν την ασθένεια. Η χρήση οντολογιών[21] αλλά και άλλων στοιχείων αποδεικνύουν ότι, παρά το θόρυβο / ελλιπή πειραματικών δεδομένων αλλά και τις ημιτελείς οντολογίες ασθενειών, ο εντοπισμός των υποψηφίων γονίδια μπορεί να είναι

επιτυχής. Ακόμη και όταν για ένα μεγάλο αριθμό υποψηφίων ασθενειών, οι όροι προβλέπονται ταυτόχρονα.

Σε αυτό τον αλγόριθμο, παρουσιάζεται μια νέα προσέγγιση για την πρόβλεψη των γονιδίων που σχετίζονται με ασθένειες με βάση πειραματικά PPIs δίκτυα, συσχετίσεις πρωτεΐνης-νόσου, καθώς και σε δεδομένα της πρωτεϊνικής ακολουθίας και λειτουργικών περιγραφών. Προτείνετε μια μέθοδος για να συνδέσει και να ταξινομήσει τα γονίδια στα διάφορα επίπεδα της νόσου ανάλογα με τις Οντολογίες της Νόσου «DO»[22]. Οργανώνει τους όρους της, σε μία ιεραρχική δομή επεκτείνοντας από τους όρους της ασθένειας στην πιο ειδική ονοματολογία της νόσου με από πάνω προς τα κάτω τρόπο. Ομοίως τα GO σε DO δεδομένα αντιπροσωπεύονται ως κατευθυνόμενα γραφήματα με βάση τις UMLS έννοιες και την διεθνή ταξινόμηση των νόσων (ICD-9)[23]

Η ιεραρχική οργάνωση της DO βάσης δεδομένων είναι επωφελείς για τους αλγόριθμους αναζήτησης υποψήφιων γονιδίων που σχετίζονται με ασθένειες γιατί χωρίζουν σε διάφορα επίπεδα (κόμβους) τις συσχετίσεις μεταξύ των ασθενειών με βάση τον σχολιασμό της νόσου, επιτρέποντας έτσι τη στατιστική συμπερασματολογία με μεγαλύτερη εμπιστοσύνη. Με την προσέγγιση αυτή θεωρούμε ότι το σύνολο του διαθέσιμου δικτύου PPIs είναι για τον άνθρωπο και κωδικοποιεί κάθε γονίδιο με βάση την κατανομή των αποστάσεων της συντομότερης διαδρομής για όλα τα γονίδια που σχετίζονται με τη νόσο ή έχουν κοινή λειτουργική περιγραφή. Επιπλέον, χρησιμοποιεί τις πληροφορίες που παίρνει από τις ιδιότητες της κάθε πρωτεϊνικής ακολουθίας. Αυτές οι πληροφορίες συνδέονται με ορισμένες κατηγορίες ασθενειών που σχετίζονται με τις πρωτεΐνες και τις ενσωματώνει μέσα από ένα πλαίσιο χρησιμοποιώντας δύο επίπεδα, με την χρήση διανυσματικών μηχανών στήριξης (SVMs). [19]

1.1.1.1 Υλικά και μέθοδοι του αλγόριθμου PhenoPred .

Έχουμε τις *GPPI, GDO, GGo, GP-GO* και *GP-DO* γραφικές παραστάσεις που αντιπροσωπεύουν τις *PPIs, DO, GO, πρωτεΐνες-GO* και *πρωτεΐνες-DO* συσχετίσεις αντίστοιχα. Ορίζουμε:

(i) *GPPI=(P, EP)*, ως ένα μη-κατευθυνόμενο γράφημα, όπου *P= (p1, p2,. . . , P|P|)* είναι ένα σύνολο από πρωτεΐνες και το $EP \leq P X P$,

(ii) *GDO=(D, ED)* ως κατευθυνόμενο άκυκλο γράφημα που αντιπροσωπεύει τις οντολογίες των ασθενειών, *όπου D= (d1, d2,. . . , D/D/)* είναι ένα σύνολο από όρους ασθενειών και το *ED< D X D;*

(iii) *GGO = (F, EF)* ως κατευθυνόμενο άκυκλο γράφημα με *GO*, όπου *F=(f1,f2,...,F/F/)* είναι ένα σύνολο λειτουργικών όρων και *EF< F X F;*

(iv) *GP-DO = (P, D, EP-DO)* ως διμερές γράφημα των συσχετίσεων πρωτεΐνης-*DO*, όπου *EP-DO (P=D),* και

(v) *GP-GO 5 (P, F, EP-GO)* ως διμερές γράφημα πρωτεΐνης-*GO* συσχετίσεις, όπου *EP-GO<P X F.*

Ο σκοπός είναι μέσα από το αλγόριθμο αυτό με την βοήθεια των γραφημάτων *GPPI, GDO, GGO, GP-GO*, και *GP-DO*, να μπορεί να προβλέψει σωστά καινούργιες συσχετίσεις μεταξύ πρωτεϊνών – ασθενειών. Κάθε σύνδεση πρωτεΐνης – ασθένειας (p, d) περιέχει μία πρωτεΐνη $p \in P$ και ένα όρο της νόσου $d \in D$. Οι όροι "πρωτεΐνη' 'και "γονίδιο" χρησιμοποιούνται κάπως εναλλακτικά, δεδομένου ότι μόνο οι πρωτεΐνες που κωδικοποιούνται από γονίδια εξετάζονται εδώ.

1.1.1.2 Σύνολα δεδομένων αλγόριθμου PhenoPred.

Οι ασθένειες και τα γονίδια των γνωστών γενετικών συμμετοχών προέρχονται από τις βάσεις δεδομένων OMIM[24], Swiss-Prot[25] και HPRD.[26] Τα ονόματα των ασθενειών και των συναφών γονιδίων συλλέχθηκαν και εντάχθηκαν στην βάση δεδομένων DO. Αποκλείστηκαν οι συσχετίσεις μεταξύ των γονιδίων και ασθενειών όπως για παράδειγμα, τα γονίδια που αποτελούν μέρος μεγάλων μετατοπισθέντων τμημάτων και κατά κανόνα συνδέονται με πολλές μορφές καρκίνου. Το δίκτυο αλληλεπίδρασης πρωτεϊνών (PPIs) δημιουργείται συνδυάζοντας τη φυσική αλληλεπίδραση των δεδομένων από τις βάσεις δεδομένων HPRD[26] και OPHID[27]. Συνολικά, ο αριθμός των πρωτεϊνών, των ασθενειών, των PPIs δικτύων, των συσχετισμένων πρωτεϊνικών λειτουργιών και των συσχετίσεων πρωτεϊνών με ασθένειες ήταν |P|=9590, |D|=14647, |EP|=41,456, |EP-GO|=235.925, |EP-DO|=55,127 αντίστοιχα. Ο συνολικός αριθμός των πρωτεϊνών που σχετίζονται με τουλάχιστον μία ασθένεια ήταν 2000, ενώ ο αριθμός των όρων που συνδέονται με τουλάχιστον μία πρωτεΐνη ήταν 2200. Τα δεδομένα αυτά είναι ελεύθερα διαθέσιμα στην ιστοσελίδα του αλγόριθμου Phenopred[13].

1.1.1.3 Αναπαράσταση δεδομένων αλγόριθμου PhenoPred.

Για κάθε πρωτεΐνη p ϵ P, κατασκευάζονται τρεις σειρές με διαφορετικά χαρακτηριστικά για την πρόβλεψη των συσχετίσεων της ασθένειας:

(i) PPI-DO, τα χαρακτηριστικά έχουν κατασκευαστεί με βάση την κατανομή της συντομότερης απόσταση από την p (πρωτεΐνη) μεταξύ των άλλων πρωτεϊνών του δικτύου αλληλεπίδρασης πρωτεϊνών PPIs. Σε αυτό το δίκτυο είναι γνωστό ότι οι πρωτεΐνες σχετίζονται με συγκεκριμένες ασθένειες.

(ii) PPI-GO, τα χαρακτηριστικά αυτά κατασκευάστηκαν με παρόμοιο τρόπο, αλλά με βάση την συντομότερη απόσταση μεταξύ των άλλων πρωτεϊνών που είναι γνωστό ότι σχετίζονται με συγκεκριμένους GO όρους.

(iii) SPP-GO χαρακτηριστικά που προέρχονται από τις διάφορες κωδικοποιημένες ακολουθίες, τις φυσικοχημικές και άλλες προβλεπόμενες ιδιότητες της πρωτεΐνης, καθώς και τους GO όρους της πρωτεΐνης.

Για την κατασκευή των PPI-DO (και αντίστοιχα PPI-GO) χαρακτηριστικών, υπολογίζονται πρώτα οι συντομότερες αποστάσεις μεταξύ όλων των ζευγών από τις πρωτεΐνες στο δίκτυο PPIs . Για κάθε συνδυασμό από P ϵ p, d ϵ D, και t ϵ (1, 2,. . . , tmax) , όπου tmax=14 (Το μέγιστο που παρατηρήθηκε με μικρότερη απόσταση σε GPPI), συν-υπολογίζονται:

(i) N_{pd}^{t} ο αριθμός των πρωτεϊνών, με την συντομότερη απόσταση από το t σε p που σχετίζονται με τη νόσο d,

(ii) N_{pd} ο αριθμός του συνόλου των πρωτεϊνών που από την p πρωτεΐνη συνδεόμαστε με την d ασθένεια, και

(iii) N_{p}^{t} ο αριθμός όλων των πρωτεϊνών με τη μικρότερη απόσταση t σε μια πρωτεΐνη.

Τα $N_{pd} = \sum t\, N_{pd}^{t}$ και $N_{pd}^{t} \leq N_{p}^{t}$ αλλά $N_{p}^{t} = \sum d\, N_{pd}^{t}$ δεν είναι αναγκαστικά συνδεδεμένα με συσχετίσεις από διάφορες ασθένειες με την ίδια πρωτεΐνη αλλά και ούτε αλληλοαναιρούνται. Τα PPI-DO χαρακτηριστικά υπολογίζονται σαν N_{pd}^{t}/N_{pd} και N_{pd}^{t}/N_{p}^{t} για κάθε d ϵ D και t ϵ (1, 2,. . . ,tmax).

Το N_{pd}^{t}/N_{pd} αντιπροσωπεύει την κατανομή της συντομότερης απόστασης από την p πρωτεΐνη σε όλες τις πρωτεΐνες που είναι γνωστό ότι σχετίζονται με τη νόσο d, ή απλά η τιμή της απόστασης με τη νόσο του d. Από την άλλη το N_{pd}^{t}/N_{p}^{τ} αναφέρει τα

κλάσματα πρωτεϊνών που συνδέονται με τη νόσο του d μεταξύ του επιπέδου t και των p γειτόνων. Υπάρχει η υπόθεση ότι μία p πρωτεΐνη που σχετίζεται με τη νόσο d είναι πιο πιθανό να μοιραστεί την κατανομή αποστάσεων στους DO όρους με τις πρωτεΐνες που συνδέονται με την ασθένεια d παρά με τις υπόλοιπες πρωτεΐνες. Στην πραγματικότητα δεν είναι όλα τα *2 X tmax/D/* χαρακτηριστικά αναγκαία, δεδομένου ότι οι πρωτεΐνες με μεγάλη τιμή σε GPPI είναι λιγότερο πιθανό να μοιραστούν DO σχολιασμούς. Έτσι, σε κάθε αναζήτηση που εκτελούμε, έχουμε συγκεντρωτικά όλα τα χαρακτηριστικά N_{pd}^{t} $/N_{pd}$ και N_{pd}^{t} $/N_{p}^{\tau}$ για $t \geq 4$, όπως $\sum_{t \geq 4} N_{pd}^{t}$ $/N_{pd}$ και $\sum_{t \geq 4} N_{pd}^{t}$ $/N_{p}^{\tau}$ αντίστοιχα.

Επιπλέον, οι DO όροι εξαιρούνται με τις λιγότερες 10 θετικές πρωτεΐνες ,από την κατασκευή των χαρακτηριστικών δεδομένων μέχρι τα προκύπτοντα χαρακτηριστικά δεδομένα να είναι λιγότερο πιθανό στατιστικά σημαντικά. Η αλληλουχία δεδομένων και τα λειτουργικά χαρακτηριστικά (SPP-GO) κατασκευάζονται με βάση:

- την πραγματική τιμή των διανυσματικών δεδομένων που λαμβάνεται για κάθε φυσικοχημικές ή προβλεπόμενες ιδιότητες.
- την δυαδική κωδικοποίηση του γνωστού GO σχολιασμού και των PROSITE[28] ελέγχων.

Η πραγματική εκτίμηση και αναπαράσταση των δεδομένων μιας πρωτεΐνης μπορεί εύκολα να λαμβάνεται με την πρόβλεψη των ιδιοτήτων της.

1.1.1.4 Μείωση διαστάσεων και εκπαίδευση μοντέλου.

Λόγω της ύπαρξης της υπέρ-τοποθέτησης και του υπολογιστικού κόστους στις ταξινομήσεις δεδομένων, η μείωση διαστάσεων είναι ένα θέμα που απασχολούσε τους δημιουργούς του αλγόριθμου. Έτσι για την μείωση τους κατατάσσονται τα διάφορα χαρακτηριστικά δεδομένα με βάση τις πληροφορίες και στη συνέχεια διατηρούνται τα χαρακτηριστικά δεδομένα με τις υψηλότερες τιμές για το K-th επαρκώς ανόμοιο χαρακτηριστικό δεδομένο, όπου K είναι ένας προκαθορισμένος αριθμός. Το χαρακτηριστικό δεδομένο Xi θεωρείται ένα αρκετά ανόμοιο επιλεγμένο χαρακτηριστικό δεδομένο που έχει ανακαλύψει ο αλγόριθμος μας από τα προηγούμενα χαρακτηριστικά δεδομένα, εάν ο μέγιστος συντελεστής ανά ζεύγος συσχέτισης μεταξύ Xi και κάθε επιλεγμένου χαρακτηριστικού δεδομένου (εκτός από τα $X_1, X_2, \dots, X_{i-1}$) ήταν κάτω από το όριο p. Η ομοιότητα μεταξύ των χαρακτηριστικών δεδομένων μετράτε με τη συσχέτιση του συντελεστή Pearson. Τέλος, συνδέει τα χαρακτηριστικά δεδομένα που

μειώνονται περαιτέρω με την εφαρμογή της ανάλυσης των κυρίων συνιστωσών και που διατηρούν σPCA διακύμανση.

Οι προβλέψεις για μεμονωμένες ασθένειες δημιουργήθηκαν με διανυσματικές μηχανές στήριξης (SVM) χρησιμοποιώντας την αρχή «μία-ενάντια-όλον». Επίσης χρησιμοποιεί τον αλγόριθμο SVMperf [29] για να βελτιστοποιηθεί η περιοχή κάτω από την χαρακτηριστική καμπύλη ROC (AUC), λόγω της ακραίας ανισορροπίας δεδομένων που χρησιμοποιούνται στην εκπαίδευση. Για κάθε πρόγνωση, καταγράφεται ο μέσος όρος και η τυπική απόκλιση των αποτελεσμάτων πρόβλεψης που σχετίζονται με την κατάρτιση δεδομένων και τις χρησιμοποιεί για να ομαλοποιήσει την τελική βαθμολογία πρόβλεψης για κάθε πρωτεΐνη με βάση τον μετασχηματισμό z-score.

Με τον τρόπο αυτό, αναμένεται ότι οι τελικές βαθμολογίες πρόβλεψης μέσα από τις διάφορες δοκιμές για όλες τις ασθένειες θα έχουν μέσο όρο περίπου κοντά στην τιμή 0 και τυπικές αποκλίσεις, τιμή περίπου στο 1. Στον αλγόριθμο αυτό κατασκευάστηκε μια ξεχωριστή ικανότητα πρόβλεψης για κάθε τύπο χαρακτηριστικών δεδομένων (PPI-DO, PPI-GO, SPP-GO) και τα συνδυάζει χρησιμοποιώντας ένα δεύτερο μοντέλο. Το δεύτερο μοντέλο έχει εκπαιδευτεί με τα ίδια δεδομένα εκπαίδευσης με την δημιουργία επιμέρους μοντέλων.

1.1.1.5 Αξιολόγηση των επιδόσεων του αλγορίθμου Phenobred.

Οι προσεγγίσεις του αλγόριθμου έχουν αξιολογηθεί 100 φορές με την μέθοδο «cross validation». Η κατασκευή των PP-DO/ PPI-GO χαρακτηριστικών δεδομένων και DO/GO συσχετίσεων στις υπό δοκιμή πρωτεΐνες όπου τους αφαιρέθηκαν και μειώθηκαν κάποια χαρακτηριστικά, έγινε με χρήση των καλά εκπαιδευμένων πρωτεϊνών. Λόγω των διαθέσιμων υπολογιστικών πόρων, δεν επιχειρήθηκε η βελτιστοποίηση των K και p τιμών για την μείωση χαρακτηριστικών και παραμέτρων στην εκπαίδευση των διανυσματικών μηχανών στήριξης, αλλά μόνο έκθεση των αποτελεσμάτων που επιτυγχάνονταν χρησιμοποιώντας K =5 και p=0.7 μαζί με τις προεπιλεγμένες παραμέτρους της μεθόδου SVM^{perf}[29]. Κατά τον ίδιο τρόπο, η σPCA διακύμανση διατηρήθηκε κατά 95%.

Για τη μέτρηση της συνολικής απόδοσης πρόβλεψης, εξετάστηκε η καμπύλη ευαισθησίας (curve of recall) ως συνάρτηση της ειδικότητας. Για κάθε υπο-δοκιμή πρωτεΐνη p, επιλέγονται οι k κορυφαίες πρωτεΐνες (k = 1 ... |D|) των προβλέψιμων όρων της νόσου και υπολογίζεται η καμπύλη ευαισθησίας ως $|D_O \cap D_P|/|D_P|$ όπου D_O είναι το

σύνολο των παρατηρούμενων ασθενειών που συνδέονται με την πρωτεΐνη και D_P είναι το σύνολο των προβλεπόμενων ασθενειών.[30] Το σημείο ευαισθησίας - ειδικότητας παρίσταται ως ο μέσος όρος για όλες τις υπο-δοκιμή πρωτεΐνες για τα k δεδομένα, όπου το δεξιότερο σημείο αντιστοιχεί με k = 1. Για κάθε τρέξιμο ένα αθροιστικό γράφημα αντιπροσωπεύει το σύνολο των όρων, του G_{DO}. Για κάθε ασθένεια, υπολογίζετε επίσης η καμπύλη ROC με την τετμημένη ευαισθησία ως συνάρτηση των ψευδώς θετικά αποτελεσμάτων. Το ποσοστό των ψευδώς θετικών υπολογίστηκε ως 1-$|\overline{D_O} \cap \overline{D_P}|$ /$|\overline{D_P}|$ όπου $\overline{D_P}$ είναι ένα συμπλήρωμα του D.

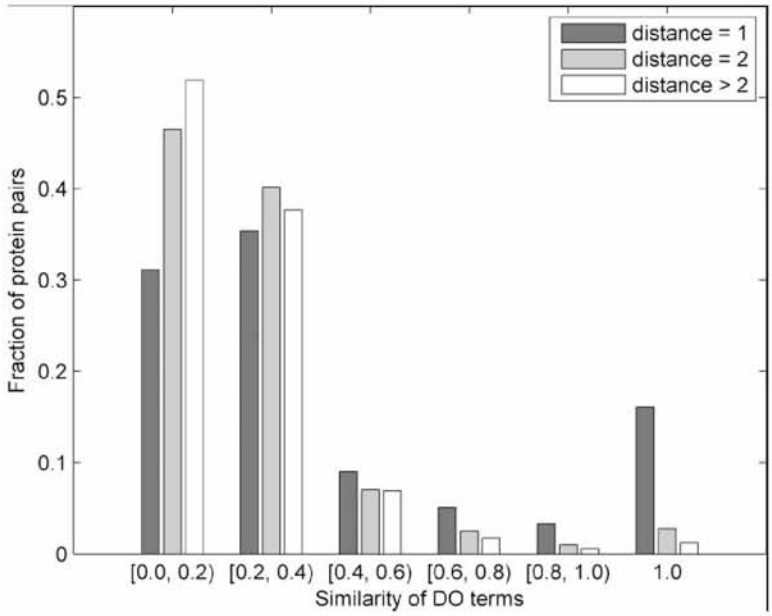
1.1.1.6 Ανάλυση του δικτύου αλληλεπίδρασης πρωτεϊνών.

Έχει ήδη παρατηρηθεί ότι οι πρωτεΐνες που αλληλεπιδρούν άμεσα είναι πιο πιθανό να μοιράζονται τους λειτουργικούς τους σχολιασμούς, και πιθανώς να σχετίζονται με ασθένειες[31]. Στο Εικόνα 1.1.1.1, αναλύεται το πρόβλημα αυτό και απεικονίζετε το κλάσμα των πρωτεϊνικών ζευγών ως συνάρτηση της ομοιότητας των μεταξύ τους όρων για κάθε νόσο. Η ομοιότητα μεταξύ των δύο όρων της κάθε νόσου υπολογίστηκε ως το κλάσμα του μέγεθος του συνόλου των κοινών όρων και το σύνολο όλων των άλλων όρων που συνδέονται με τους όρους της νόσου. Τα κλάσματα παρουσιάζονται χωριστά για τις άμεσα αλληλεπιδρόμενες πρωτεΐνες, για τα ζεύγη πρωτεϊνών σε απόσταση 2 και για εκείνες με απόσταση μεγαλύτερη από 2. Μόνο περίπου το 16% των πρωτεϊνών που αλληλεπιδρούν άμεσα έχουν από κοινού ακριβή DO σχολιασμό, περίπου το 17% των ζευγαριών έχει από κοινού κάπως παρόμοιους σχολιασμός (με ομοιότητα μεταξύ 0,4 και 1), ενώ πάνω από το 66% στα απευθείας αλληλεπιδρόμενα ζεύγη έχουν πολύ διαφορετικούς σχολιασμούς.

Έτσι, για έναν αλγόριθμο που σχεδιάστηκε να προβλέπει συσχετίσεις μεταξύ γονιδίων – νόσου, είναι σημαντικό να συμπεριλάβει πληροφορίες αναφορικά με την απόσταση / ποσοστό σύνδεσης, αλλά και τις πιο απομακρυσμένες γειτονικές περιοχές (αυτή είναι η λογική πίσω από την αναπαράσταση δεδομένων στα τμήματα των συνόλων δεδομένων). Το κλάσμα των ζευγών στο Εικόνα 1.1.1.1 έχει ληφθεί με βάση την αναγωγή του αριθμού των πρωτεϊνών σε απόσταση 1, 2, και >2 ξεχωριστά από το σύνολο σε κάθε μία από αυτές τις κατηγορίες (2519, 46.306 και 1.101.061 αντίστοιχα). Για καλύτερη επεξήγηση, το 16% των άμεσων γειτονικών περιοχών που μοιράζονται το ίδιο σχόλιο νόσου αντιστοιχούν σε 405 ζεύγη (405/2519 = 0.16), ενώ το 3% των ζευγών

των πρωτεϊνών σε απόσταση 2 όπου μοιράζετε τον ίδιο σχολιασμό αντιστοιχεί σε 1293 ζευγάρια (Εικόνα . 1.1.1.1).

Εικόνα 1.1.1.1: Κατανομή των κλασμάτων από τα διάφορα ζεύγη πρωτεϊνών με απόσταση $d \in (1,2,>2)$ στο ανθρώπινο δίκτυο αλληλεπίδρασης PPIs με την ομοιότητα συσχετισμού των DO όρων τους με την ασθένεια.



1.1.1.7 Προβλεπτική ακρίβεια του αλγόριθμου PhenoPred.

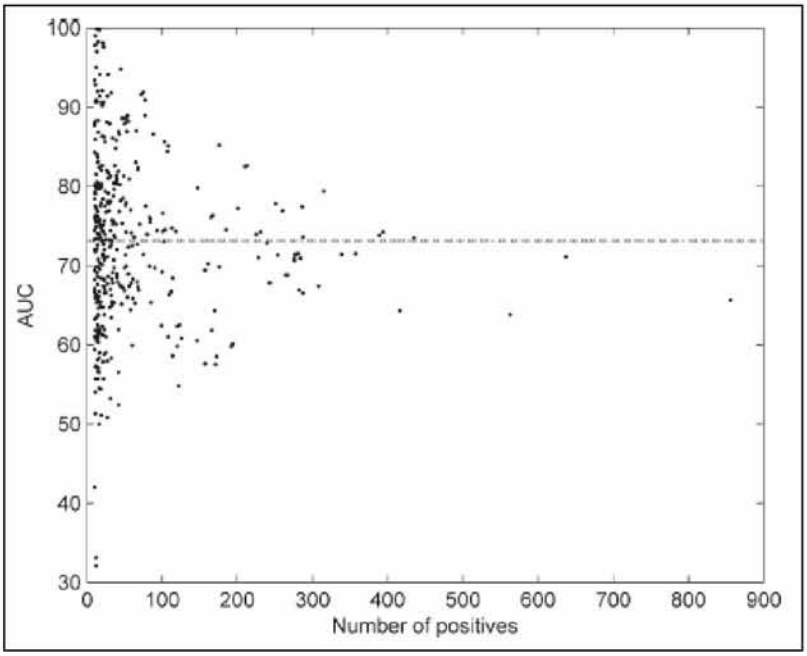
Η απόδοση των ταξινομητών αξιολογήθηκε ξεχωριστά για μεμονωμένες ασθένειες και, για τις συνολικές επιδόσεις του ολοκληρωμένου μοντέλου. Από τη στιγμή που χρησιμοποιείται το δίκτυο αλληλεπίδρασης πρωτεϊνών PPIs για την κατασκευή των χαρακτηριστικών δεδομένων PPI-DO και PPI-GO, λόγω των ανεπαρκή G_{PPI} δεδομένων, υπάρχει μια σειρά από μικρά συνδεδεμένα αθροιστικά γραφήματα, τα οποία εμπόδισαν τους σχεδιαστές στην κατάρτιση ενός μοντέλου χρησιμοποιώντας όλες τις διαθέσιμες πρωτεΐνες. Αντί αυτού, το μοντέλο ταξινόμησης έχει εκπαιδευτεί για τα πολύ συνδεδεμένα αθροιστικά γραφήματα του G_{PPI}, που περιέχουν 8934 κόμβους, ή 93% του συνολικού αριθμού των πρωτεϊνών. Από αυτούς, 1517 συσχετίζονται με την νόσο.

Παρά το γεγονός ότι για κάθε ασθένεια το σύνολο των θετικών όρων περιέχονται στα γονίδια που συνδέονται με τον συγκεκριμένο όρο της νόσου, το σύνολο των αρνητικών περιείχε όλα τα υπόλοιπα που σχετίζονται με άλλες ασθένειες και γονίδια, καθώς και το 10% των μη συνδεδεμένων με ασθένειες γονιδίων επιλέγονται τυχαία (λόγω του μεγάλου αριθμού). Τα προβλεπτικά μοντέλα έχουν εκπαιδευτεί μόνο για

όρους της νόσου που έχουν 10 ή περισσότερα γονίδια που να συνδέονται μαζί τους. Αν δύο ή περισσότεροι όροι από ασθένειες είχαν όμοια σύνολα συνδεδεμένων γονιδίων, τότε διατηρούνται μόνο οι πιο συγκεκριμένοι όροι. Με βάση την διαδικασία, ο αλγόριθμος αυτός χρησιμοποιεί 422 όρους που σχετίζονται με ασθένειες με τους οποίους έγινα οι προβλέψεις.

Η Εικόνα 1.1.1.2 παρουσιάζει την περιοχή κάτω από την καμπύλη ROC (AUC) για όλους τους 422 επιμέρους όρους που σχετίζονται με τις νόσους σε συνάρτηση με το μέγεθος της εκπαίδευσης που έχουν τεθεί. Η μέση AUC εκτιμήθηκε σε 73,1%, και μπορεί να παρατηρηθεί ότι η ειδικότητα μειώνεται ελαφρώς με την αύξηση του αριθμού των θετικών όρων.

Εικόνα 1.1.1.2: Περιοχή κάτω από την καμπύλη ROC (AUV) σε συνάρτηση των θετικών παραδειγμάτων για τους 422 όρους που σχετίζονται με ασθένειες.

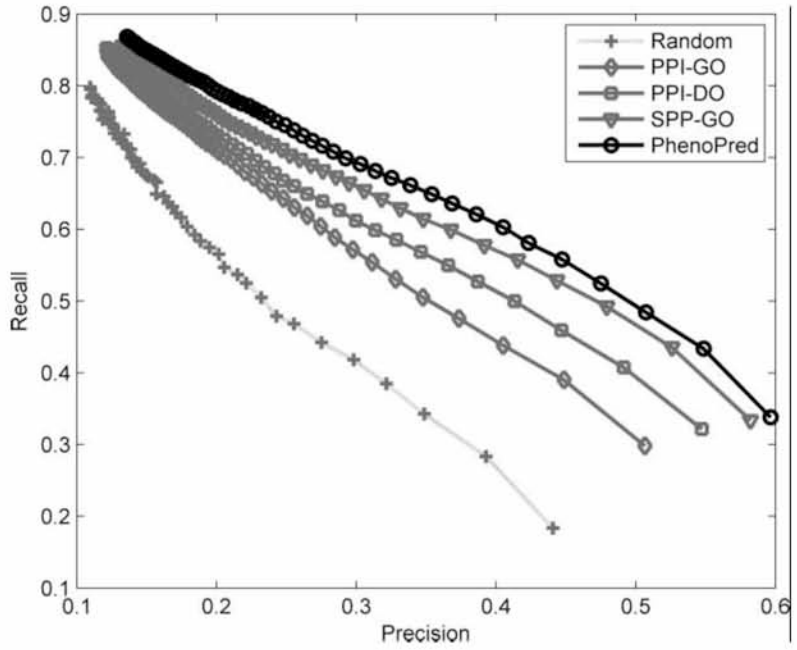


Αυτό είναι αναμενόμενο, δεδομένου ότι όσο ανεβαίνουμε πιο κοντά στη ρίζα του DO, η παθήσεις γίνονται όλο και πιο γενικές και πιο δύσκολο να προβλεφθούν.

Η Εικόνα 1.1.1.3 δείχνει την ευαισθησία ως συνάρτηση της ειδικότητας για τους τρεις επιμέρους ταξινομητές στο συνδυασμένο μοντέλο (PhenoPred), το οποίο λαμβάνεται με την κατάρτιση του μοντέλου δεύτερου σταδίου. Είναι σημαντικό, και τα τρία είδη δεδομένων που εκτελούνται ναι είναι από σωστή επιλογή και όχι από τύχη, ωστόσο, οι διαφορές μεταξύ των ταξινομητών ήταν σημαντικές. Ο ταξινομητής με βάση

τα PPI δίκτυα και των GO συσχετίσεων (PPI-GO) είχε τις πιο αδύναμες επιδώσεις. Μετά ακολουθούσε ο ταξινομητής με βάση τα PPI δίκτυα και των DO συσχετίσεων (PPI-DO), και τέλος ο ταξινομητής με βάση την αλληλουχία και τις GO πληροφορίες (SPP-GO) είχε τη μεγαλύτερη ακρίβεια. Ο συνδυασμένος ταξινομητής επιτύγχανε τις καλύτερες επιδόσεις με κατά μέσον όρο, περίπου το 60% να προβλέπει σωστά τους όρους. Σε επίπεδο ακρίβειας της τάξης του 45%, ο αλγόριθμος PhenoPred έχει 40 εκατοστιαίες μονάδες υψηλότερο από ό, τι το τυχαίο μοντέλο.

Εικόνα 1.1.1.3: Ευαισθησία ως συνάρτηση της ειδικότητας για τους 3 επιμέρους ταξινομητές και τον ταξινομητή Phenobred.



1.1.2 Αλγόριθμος Gene Prospector.

Ένα στοιχείο του τεράστιου πλοηγού, μια ολοκληρωμένη αναζήτηση γνώσεων για τις γενετικές ενώσεις και τις σχετικές πληροφορίες στην ανθρώπινη γονιδιακή επιδημιολογία. Ο Gene Prospector[32] είναι ένα βιοπληροφορικό εργαλείο που σχεδιάστηκε για να ταξινομεί και να κατατάσσει σε ιεραρχική σειρά γονίδια αλλά και τις σχετικές τους πληροφορίες για τις ανθρώπινες γονιδιακές επιδημιολογίες με βάση παράγοντες της ασθένειας και άλλων φαινότυπων. Είναι μία αποδεικτική πύλη στοιχείων για την εκτίμηση των πιθανών ευπαθή γονιδίων και την αλληλεπίδραση των παραγόντων κινδύνου για τις ανθρώπινες ασθένειες.

Η διαδικτυακή αυτή εφαρμογή επιλέγει και ιεραρχεί σε σειρά ενδεχόμενα (πιθανά) γονίδια που σχετίζονται με ασθένειες χρησιμοποιώντας μια πολύ καλά επιλεγμένη και συνεχόμενα ενημερωμένη βάση δεδομένων από διεθνή μελέτες για το γενετικό κώδικα. Επίσης παρέχει ένα ολοκληρωμένο σύνολο στις αναζητήσεις με βάση τις λέξεις κλειδιά σε συμπληρωματικές πηγές δεδομένων.

Ως αποτέλεσμα της εύρεσης του ανθρώπινου γονιδιώματος[33] αλλά και βελτίωσης της τεχνολογίας, έχουμε μία μεγάλη αύξηση των γενετικών μελετών. Πρόσφατα οι μελέτες πάνω στο ευρύ φάσμα του γονιδιώματος έχουν αρχίσει να εξετάζουν σημαντικά μεγάλο αριθμό γενετικών συναφειών[34]. Η σύνθεση των εν λόγω πληροφοριών είναι το πρώτο βήμα στην μετάφραση των νέων γνώσεων που αναπτύχθηκαν από την βασική έρευνα για τις κλινικές εφαρμογές στην δημόσια υγεία[35].

Παρά το γεγονός ότι αρκετές βάσεις δεδομένων με γονίδια και ασθένειες διατίθενται δωρεάν, η εύρεση δημοσιευμένων ερευνών που να σχετίζουν γονίδια με ασθένειες και η κατανόησης των σχέσεων της νόσου αλλά και των αλληλεπιδράσεων των γονιδίων με το περιβάλλον δεν είναι ασήμαντη εργασία. Ο αλγόριθμος Gene Propspector είναι μία διαδικτυακή εφαρμογή σχεδιασμένη να ιεραρχεί και να αξιολογεί στοιχεία για τα γονίδια που σχετίζονται με τις ανθρώπινες ασθένειες ή τις αλληλεπιδράσεις με γενετικούς παράγοντες κινδύνου. Ο αλγόριθμος αυτός χρησιμοποιεί τα υποψήφια γονίδια που βρίσκει με βάση την δημοσιευμένη βιβλιογραφία[36] από την βάση δεδομένων Huge, αλλά και κάνει γρήγορες αναζητήσεις σε όλες τις πηγές δεδομένων .

Ταξινομεί τα διάφορα γονίδια που βρίσκει να σχετίζονται με βάση τις δημοσιεύσεις αλλά και τα βάζει σε σειρά ανάλογα με τον αριθμό των δημοσιεύσεων που

βρίσκει για το κάθε γονίδιο στην εν λόγο συσχέτιση με την ασθένεια. Επιπλέον στην έρευνα του προσθέτει τα γονίδια που βρέθηκαν να σχετίζονται με την ασθένεια που αναζητούμε μέσω δημοσιεύσεων για δύο ζώα τον αρουραίο και τον ποντικό. Οι αναζητήσεις γίνονται με βάση τα Mesh terms. Τα Mesh terms είναι ιατρικοί όροι που προσκαλούνται στις καταχωρίσεις του κάθε άρθρου στην HUGE βάση δεδομένων και αυτοί οι όροι δίνονται με βάση την NCBI[37] βάση δεδομένων. Η χρήση αυτών των Mesh όρων γίνεται για την γρήγορη και αποτελεσματική αναζήτηση, με βάση τις δενδρικές αναζητήσεις.

1.1.2.1 Επιλογή και ιεράρχηση των γονιδίων.

Για την επιλογή και ιεράρχηση των γονιδίων δημιουργείται ένας κατάλογος αναζήτησης με βάση τις δημοσιεύσεις που βρέθηκαν. Για κάθε γονίδιο βρίσκουμε τον αριθμό των δημοσιεύσεων σε διαφορετικές κατηγορίες (συνολικά, για τις γενετικές συνδέσεις, ενώσεις με βάση το ευρύ γονιδίωμα, με βάση τα αποτελέσματα συγκεντρωτικών μετά-αναλύσεων) που εμφανίζονται. Ταξινομούνται σε ιεραρχική σειρά με βάση την συνολική βαθμολογία (score) που συγκεντρώνουν:

$$Score = \frac{Hi}{\sum_{i=1}^{n} Hi} + \frac{GAi}{\sum_{i=1}^{n} GAi} + \frac{GWASi}{\sum_{i=1}^{n} GWASi} + \frac{MAi}{\sum_{i=1}^{n} MAi} + \frac{GTi}{\sum_{i=1}^{n} GTi}$$

- Hi: Αριθμός του συνόλου των δημοσιεύσεων για ένα συγκεκριμένο γονίδιο και ο χρόνος αναζήτησης.
- Gai: Αριθμός δημοσιεύσεων σε γενετικές μελέτες σύνδεσης για ένα συγκεκριμένο γονίδιο και ο χρόνος αναζήτησης.
- GWASi: Αριθμός εκδόσεων σε επίπεδο μελετών genome_wide δημοσιεύσεων και χρόνο αναζήτησης.
- Mai: Αριθμός δημοσιευμένων μετά-αναλύσεων για το δοσμένο γονίδιο και χρόνος αναζήτηση.
- GTi: Αριθμός των δημοσιευμένων γενετικών πειραμάτων για ένα δεδομένο γονίδιο και χρόνος αναζήτησης.

Αν κατά την ιεράρχηση δύο συνολικών βαθμολογιών έχουμε την ίδια τιμή τότε προτεραιότητα στην ιεράρχηση παίρνει το γονίδιο για το οποίο έχουν βρεθεί πειραματικές δοκιμές που βασίζονται σε ζώα. Σε κάθε αναζήτηση παίρνουμε

πληροφορίες για κάθε γονίδιο από την SNPs βάση δεδομένων για συνώνυμα, μη συνώνυμα, splice sites, UTR-unntraslated region και intron.

1.1.2.2 Αποτελέσματα αλγόριθμου GeneProspector.

Ο Gene Prospector με μεγάλη αξιοπιστία βρίσκει τα υποψήφια γονίδια που σχετίζονται με τις ασθένειες, με τις μεθόδους που προαναφέρθηκαν αλλά και τα ταξινομεί με βάση την συνολική βαθμολογία (score). Με βάση έρευνες που έγιναν για το κατά πόσο ο αλγόριθμος αυτός ανιχνεύει σε μεγάλο ποσοστό σωστά γονίδια, με βάση λίστες γονιδίων που σχετίζονται με ασθένειες που έχουν βρεθεί από πειραματικά δεδομένα, ο Gene Prospector κατά 75% ανίχνευσε ορθά τα ήδη γνωστά γονίδια και παρουσίασε και κάποια άλλα που πιθανός ακόμη δεν έχουν συσχετιστεί.

Σε μία ενδεικτική αναζήτηση ο αλγόριθμος αυτός συλλέγει και παρουσιάζει διαφορετικού τύπου πληροφορίες από διαφορετικές ανεξάρτητες βάσεις δεδομένων. Παρέχει συνδέσμους όπου ο χρήστης μπορεί να ανατρέξει και να δει τις δημοσιεύσεις από τις βάσεις δεδομένων HUGE[38], PUBMED[39], GWAS[34], βάσεις δεδομένων γενετικών συσχετίσεων, βάσεις δεδομένων από αποτελέσματα μετά-ανάλυσεις και γενετικών πειραμάτων και τέλος πληροφορίες από την SNP βάση δεδομένων. Στην συνέχεια στον πίνακα 1.1.2.1 παρουσιάζονται όλες οι πηγές από τις οποίες συλλέγει δεδομένα και πληροφορίες ο Gene Prospector.

Πίνακας 1.1.2.1: Πηγές αναζήτησης δεδομένων του αλγόριθμου Gene Prospector.

Όνομα	Σελίδα
Εύρεση Γονιδίων	
Enter Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Ensembl Human	http://www.ensembl.org/Homo_sapiens/index.html
Swiss-Prot	http://ca.expasy.org/sprot/
AceView	http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html?human
HuGE Navigator	http://www.hugenavigator.net/HuGENavigator/startPagePedia.do
OMIM	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
GeneCards®	http://www.genecards.org/index.shtml
Genetics Home Reference	http://ghr.nlm.nih.gov/BrowseGenes
SOURCE	http://source.stanford.edu/cgi-bin/source/sourceSearch
PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez/
Δημοσιεύσεις	
HuGE Navigator	http://www.hugenavigator.net/HuGENavigator/startPagePedia.do
Genetic Association Database	http://geneticassociationdb.nih.gov/
Φαρμακολογία	
PharmGKB	http://www.pharmgkb.org/index.jsp
Παραλλαγές/Διαδοση ασθενειών	
dbSNP	http://www.ncbi.nlm.nih.gov/sites/entrez
dbSNP-Genotype	http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi?
dbSNP-GeneView	http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=
ALFRED	http://alfred.med.yale.edu/alfred/index.asp
SNPper	http://snpper.chip.org/bio/snpper-enter-gene
Human Gene Mutation Database	http://www.hgmd.cf.ac.uk/ac/index.php
International HapMap Project	http://snp.cshl.org/index.html
The Cancer Genome Anatomy Project	http://cgap.nci.nih.gov/
Pathway	
Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/genes.html
BioCarta	http://www.biocarta.com/genes/index.asp
Pathway Interaction Database	http://pid.nci.nih.gov/PID/index.shtml
Microarrays	
NCBI Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Search&db=geo&term
Miscellaneous	
NCBI Bookshelf	http://www.ncbi.nlm.nih.gov/sites/entrez?db=books
NCBI Gene Ontology Database	http://www.geneontology.org/
GeneTests	http://www.geneclinics.org

1.1.3 Αλγόριθμος Prospectr.

Μια από της μεγαλύτερες τάσης της εποχής τα τελευταία χρόνια είναι η αύξηση του ενδιαφέροντος για προσδιορισμό των γενετικών μελετών που συχνά μελετούν και εξετάζουν μεγάλα μακρομόρια που περιέχουν εκατοντάδες γονίδια. Σκοπός του αλγόριθμου Prospectr [40] (ιεράρχηση με βάση την ακολουθία και τα φυλογενετικά πρότυπα των υποψήφιων περιοχών και γονιδίων) είναι να προβλέψει και να ιεραρχήσει τα υποψήφια γονίδια που σχετίζονται με ασθένειες αναζητώντας σε διάφορες βάσεις δεδομένων τα γονίδια που σχετίζονται με μια ασθένεια και στην συνέχεια ιεράρχηση τους με βάση τον λειτουργικό σχολιασμό και τον φαινότυπο που έχει η ασθένεια αλλά και τα υποψήφια γονίδια. Τα γονίδια που σχετίζονται με ασθένειες μοιράζονται πρότυπα χαρακτηριστικών ακολουθίας τα οποία μπορούν να μας προσφέρουν μια καλή και αξιόπιστη βάση δεδομένων για δημιουργία μία αυτοματοποιημένης μηχανικής μάθησης και ανάλυσης προτεραιότητας υποψήφιων γονιδίων.

Γενικά στον αλγόριθμο Prospectr εξετάστηκε ένα σύνολο από χαρακτηριστικά γνωρίσματα ακολουθίας και διαπιστώθηκε ότι πολλά από αυτά έχουν σημαντικές διαφορές μεταξύ του συνόλου των γονιδίων που είναι γνωστό ότι εμπλέκονται σε ανθρώπινες κληρονομικές ασθένειες από εκείνα τα σύνολα που δεν είναι γνωστό ότι εμπλέκονται με καμία ασθένεια. Ο αλγόριθμος Prospectr είναι ένας αυτόματος ταξινομητής που λειτουργεί με βάση τα χαρακτηριστικά της ακολουθίας των υποψήφιων γονιδίων χρησιμοποιώντας τον ταξινομητή δένδρου απόφασης ο οποίος κατατάσσει τα υποψήφια γονίδια με σειρά πιθανότητας συμμετοχής τους στη νόσο.

Η παραδοσιακή γονιδιακή προσεγγίζει τον τρόπο μείωσης του αριθμού των γονιδίων που είναι υπεύθυνα για την εκδήλωση κάποιων γνωρισμάτων, με τον διαχωρισμό τους σε διάφορα επίπεδα, όπου είναι μία προσπάθεια να τα ομαδοποιήσει ανάλογα με τα λειτουργικά τους σχόλια, σε γνώση της νόσου ή της φαινοτυπικής τους διερεύνησης. Δυστυχώς, αυτή η προσέγγιση έχει χαρακτηριστεί από αβάσιμους και μη αποτελεσματικούς ισχυρισμούς. Προβλήματα εγείρονται επειδή ο γονότυπος έχει σχεδόν μηδενική σχέση με το φαινότυπο σε ασθένειες πολυπαραγοντικής αιτιολογίας και το ταίριασμα των λειτουργικών σχολιασμών για ένα μονό γονίδιο με το φαινότυπο του είναι απίθανο να είναι επιτυχείς εάν δεν συνδέεται σαφώς με κάποια γνωστή παθογένεση της νόσου. Επίσης ο λειτουργικός σχολιασμός του ανθρώπου γονιδιώματος δεν είναι ακόμη ολοκληρωμένος. Η ανάθεση λειτουργικών σχόλιων είναι μια χρονοβόρα διαδικασία η

οποία είναι αναπόφευκτα επιρρεπής σε λάθη. Λαμβάνονται διάφορα μέτρα στην αντιμετώπιση των διαφόρων προβλημάτων αλλά πολλές φορές οι λειτουργικοί σχολιασμοί τους ποικίλουν ανάλογα με την βάση δεδομένων που αναζητούμε και αυτό μπορεί να παραπλανήσει ή να καθυστέρηση τους ερευνητές[41].

Έχει διατυπωθεί η άποψη ότι τα γονίδια που διέπουν τις ανθρώπινες κληρονομικές ασθένειες μοιράζονται ορισμένα χαρακτηριστικά, όπως τα χαρακτηριστικά ακολουθίας που στην περίπτωση μας, έχουν μεγαλύτερο μέγεθος τα γονίδια που σχετίζονται με ασθένειες σε αντίθεση με τα φυσιολογικά. [42] Χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης που έχουν ως στόχο να ανακαλύψουν τέτοια κοινά πρότυπα θα μπορούσαν να εφαρμοστούν για να δημιουργήσουν ένα αυτόματο σύστημα ταξινόμησης, που θα είναι σε θέση να εντοπίσει τα γονίδια τα οποία έχουν περισσότερες πιθανότητες να μην εμπλέκονται σε νόσους. Η μηχανική μάθηση προχώρησε ταχύτατα με την βοήθεια της πειραματικής τεχνητής νοημοσύνης. Οι έρευνες της Βιοπληροφορικής υιοθέτησαν αλγόριθμους μηχανικής μάθησης σε μια ποικιλία διαφορετικών καταστάσεων και η χρήση τους είναι πλέον διαδεδομένη [43]. Η προσέγγισή του αλγόριθμου Prospectr συνδυάζει και εξετάζει ένα ευρύτερο σύνολο αλγορίθμων που παράγουν σημαντικά ένα πιο επιτυχημένο ταξινομητή που είναι σε θέση να προβλέψει τα γονίδια που εμπλέκονται σε δύο κατηγορίες, τα μεντελικά και τα πιο πολύπλοκα χαρακτηριστικά.

1.1.3.1 Δημιουργία εκπαιδευμένων συνόλων και ειδικά χαρακτηριστικά.

Τα διάφορα σύνολα με τα χαρακτηριστικά που αναζητούνται για κάθε γονίδιο από τον αλγόριθμο Prospectr επιλέχθηκαν από 18.000 περίπου γνωστά γονίδια που δεν είναι γνωστό να εμπλέκονται σε ανθρώπινες ασθένειες από την Ensembl βάση δεδομένων [44] και από τα 1084 γονίδια που είναι καταχωρημένα στην Ensembl βάση δεδομένων και απαριθμούνται στην Online Mendelian Inheritance in Man (OMIM) [24] βάση δεδομένων. Το σύνολο λειτουργιών (Πίνακας 1.1.3.1) αντικατοπτρίζει τη δομή, τις λέξεις κλειδιά και το βαθμό φυλογενετικής ανάλυσης που εξετάζεται με κάθε γονίδιο (το βαθμό στον οποίο ένα γονίδιο διαφυλάσσετε πίσω μέσω της εξέλιξης και βασίζεται σε ομόλογά από άλλα είδη). Συμπεριλαμβάνονται οι σηματοδοτικές ακολουθίες και οι προβλέψεις του διαμεμβρανικού τομέα, αν και αυτά τα δεδομένα με συνδυασμό άλλων

αλγορίθμων μπορούν να υπολογιστούν με υψηλό βαθμό ακρίβειας απευθείας από την ακολουθία.

Πίνακας 1.1.3.1: Σύνολα δεδομένων που χρησιμοποιεί ο αλγόριθμός από διάφορες βάσεις δεδομένων. Όλα αυτά τα δεδομένα χρησιμοποιούνται για την δημιουργία των προφίλ για το κάθε γονίδιο αλλά και των δένδρων απόφασης.

Χαρακτηριστικά	Βάση Δεδομένων	Περιγραφή
Gene length	EnsemblMart 22.1	Μήκος γονιδίου σε bp.
CDS length	EnsemblMart 22.1	Μήκος κωδικοποιημένης ακολουθίας σε bp.
cDNA length	EnsemblMart 22.1	Μήκος ολοκληρωμένου DNA σε bp.
Protein length	EnsemblMart 22.1	Μήκος πρωτεΐνης σε aa.
Length of 3 UTR	EnsemblMart 22.1	Μήκος της με φορά 3 αποκωδικοποιημένης αλυσίδας.
Length of 5 UTR	EnsemblMart 22.1	Μήκος της με φορά 5 αποκωδικοποιημένης αλυσίδας.
Απόσταση κοντινότερου γονιδίου	EnsemblMart 22.1	Απόσταση το επόμενου γνωστού γονιδίου που βρίσκεται στο ίδιο χρωμόσωμα και με το ίδιο μήκος σε bp.
Αριθμός Εξονίων	EnsemblMart 22.1	Αριθμός εξονίων στο γονίδιο
GC	EnsemblMart 22.1	GC αζωτούχες βάσεις στο γονίδιο .
Διαμεμβρανικά	EnsemblMart 22.1	Προβλέψεις διαμεμβρανικών περιοχών .
Σηματοδοτηκά πεπτίδια	EnsemblMart 22.1	Προβλέψεις σηματοδοτικών περιοχών.
Παράλογα	EnsemblMart 22.1	Αν το ίδιο γονίδιο έχει παράλογα στο ανθρώπινο γονιδίωμα.
Αναγνωρισμένα Παράλογα	EnsemblMart 22.1	Αριθμός αναγνωρισμένων πρωτεϊνών από τα καλύτερα παράλογα στο ανθρώπινο γονιδίωμα.
Αναγνωρισμένα ομόλογα γονίδια από ποντικό	Homologene	Αριθμός αναγνωρισμένων προτείνων από ομόλογα ποντικού.
Αναγνωρισμένα ομόλογα από αρουραίο	Homologene	Αριθμός αναγνωρισμένων προτείνων από ομόλογα ποντικού.
Αναγνωρισμένα ομόλογα από σκουλήκι	Homologene	Αριθμός αναγνωρισμένων προτείνων από ομόλογα σκουληκίου .
Αναγνωρισμένα ομόλογα από μύγα	Homologene	Αριθμός αναγνωρισμένων προτείνων από ομόλογα μύγας.
Ομόλογα Ka ποντικού	Homologene	Άθροισμα από μη-συνόνημων αλλαγών μεταξύ ομόλογων γονιδίων ανθρώπου και ποντικού.
Ομόλογα Ks ποντικού	Homologene	Άθροισμα από συνόνημων αλλαγών μεταξύ ομόλογων γονιδίων ανθρώπου και ποντικού.
Ομόλογα Ka/Ks ποντικού	Homologene	Ο λόγος τον δύο παρά πάνω τομέων.
CpG island at 3 and of gene	EnsemblMart 22.1	Αν υπάρχουν CpG περιοχές στον τέλος της αλυσίδας με φορά 3.
CpG islan at 5 end of gene	EnsemblMart 22.1	Αν υπάρχουν CpG περιοχές στον τέλος της αλυσίδας με φορά 3.

Μέσα από ένα λεπτομερή έλεγχο και σύγκριση των χαρακτηριστικών συνόλων ελέγχου και των συνόλων γονιδίων που σχετίζονται με ασθένειες από δεδομένα που προήλθαν από τις βάσης δεδομένων Ensembl και OMIM βρέθηκαν να έχουν σημαντική διαφορά τα δύο σύνολα δεδομένων.[42] (Πίνακας 1.1.3.2.) Χρησιμοποιώντας το Mann-Whitney U έλεγχο βρήκαν πολύ σημαντικές διαφορές μεταξύ των γονιδίων, πρωτεϊνών

και cDNA μεγεθών από τα δύο σύνολα (P <0,001). Τα γονίδια που ανήκουν στην OMIM βάση δεδομένων ήταν σημαντικά μεγαλύτερα και κωδικοποιημένα με μεγαλύτερες πρωτεΐνες

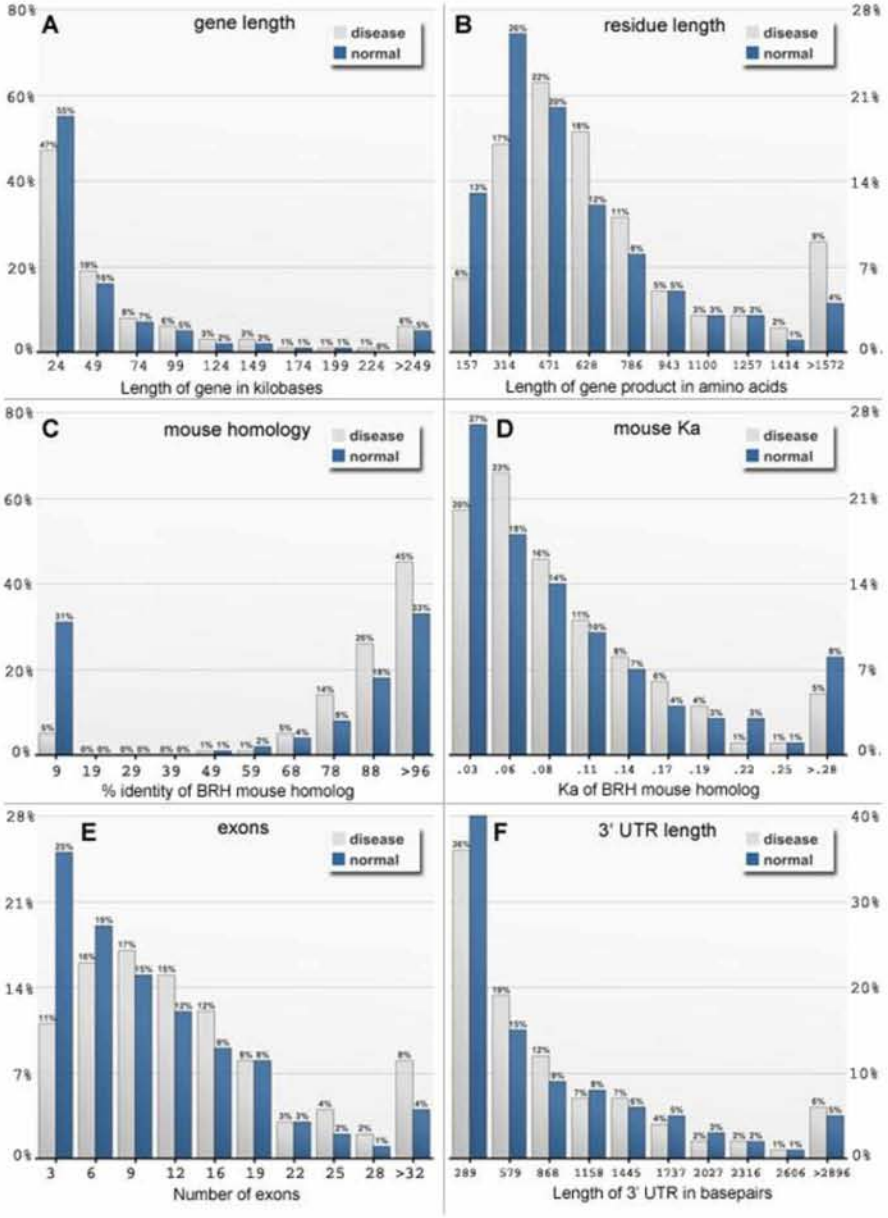
Πίνακας 1.1.3.2: Σύγκριση χαρακτηριστικών που προήλθαν από τα σύνολα γονιδίων ελέγχου έναντι χαρακτηριστικών γονιδίων που σχετίζονται με ασθένειες.

Χαρακτηριστικά	Μέσος όρος σε σύνολο γονιδίων ελέγχου	Μέσος όρος σε σύνολο γονιδίων που σχετίζονται με ασθένειες	Σημαντικότητα
Μέγεθος γονιδίου	19k	27k	P<0.001
cDNA μέγεθος	2,126bp	2.442bp	P<0.001
Μέγεθος πρωτεΐνης	383aa	494aa	P<0.001
Μέγεθος 3' UTR	446bp	488bp	P<0.01
Αριθμός εξονίων	8	10	P<0.001
Απόσταση από γειτονικό γονίδιο	46kp	52kb	P<0.01
Αναγνώριση πρωτεΐνης σε πειραματικά δεδομένα ποντικού (BRH)	80%	87%	P<0.001
Σηματοδοτημένη κωδικοποιημένη ακολουθία γονιδίου	17%	35%	P<0.001
5' CpG islands	12%	16%	P<0.028

Τα γονίδια και οι πρωτεΐνες που εμπλέκονται με ανθρώπινες ασθένειες τείνουν να είναι μεγαλύτερα από τον μέσο όρο των άλλων γονιδίων / πρωτεϊνών . Ομοίως, διαπιστώθηκε ότι τα γονίδια που απαριθμούνται στην OMIM βάση δεδομένων ήταν πολύ πιο πιθανό να έχουν καλά διατηρημένα ομόλογά με άλλα είδη και ιδίως με ποντίκια. Λόγω του μεγαλύτερου μεγέθους που έχουν τα γονίδια που εμπλέκονται με ασθένειες και της συσχέτισης μεταξύ του μεγέθους του γονιδίου και του αριθμού των εξονίων, βρέθηκε μια πολύ σημαντική διαφορά στον αριθμό των εξονίων ανά γονίδιο (P <0,001 με την χρήση του Mann - Whitney U test). Τα γονίδια που απαριθμούνται στην βάση δεδομένων OMIM είχε κατά μέσο όρο 10 εξόνια ενώ τα γονίδια που δεν είναι γνωστό ότι εμπλέκονται με ασθένειες έχουν περίπου 8. Τα γονίδια που περιλαμβάνονται στην βάση δεδομένων OMIM ήταν συχνότερα εκφρασμένα σε συγκεκριμένους τύπους ιστών (P <0,001) [45] ωστόσο, ο αλγόριθμος Prospect αποφάσισε να εξαιρεθούν τα χαρακτηριστικά δεδομένα των ιστών από το σύνολο των χαρακτηριστικών δεδομένων

που αναζητούνται για το κάθε γονίδιο , προκειμένου να αποφευχθεί η πιθανή μεροληψία των δεδομένων.

Εικόνα 1.1.3.1: Διαγράμματα με τις διάφορες διανομές των επιλεγμένων χαρακτηριστικών δεδομένων για κάθε γονίδιο.



Στα διαγράμματα (Εικόνα 1.1.3.1) παρουσιάζονται οι διάφορες διανομές των επιλεγμένων χαρακτηριστικών δεδομένων στα δύο σύνολα . Αν και ορισμένες από τις διαφορές που βρέθηκαν για κάθε γονίδιο, η ασυμφωνία για το μήκος του με φορά 3 'UTR δεν έχει τις αναμενόμενες περιπτώσεις που είχαν προσέξει και μελετήσει οι ερευνητές και δεν μπορεί να εξηγηθεί εύκολα η συσχέτιση τους. Δύο άλλα καινούργια χαρακτηριστικά είναι η απόσταση από το πλησιέστερο γειτονικό γονίδιο και ο αριθμός

των εξονίων. Και τα δύο είναι πολύ στενά συνδεδεμένα με το μέγεθος του γονιδίου (με συντελεστές συσχέτισης Spearman 0,69 και 0,71, αντίστοιχα). Μελετήθηκε επίσης ο αριθμό των πρωτεϊνικών περιοχών σε κάθε σύνολο (γονιδίων ελέγχου/ γονιδίων που σχετίζονται με ασθένειες) και διαπιστώθηκαν σημαντικές διαφορές, αλλά κατέληξαν στο συμπέρασμα ότι υπήρχε κλήση προς τα καλύτερα μελετημένα γονίδια.

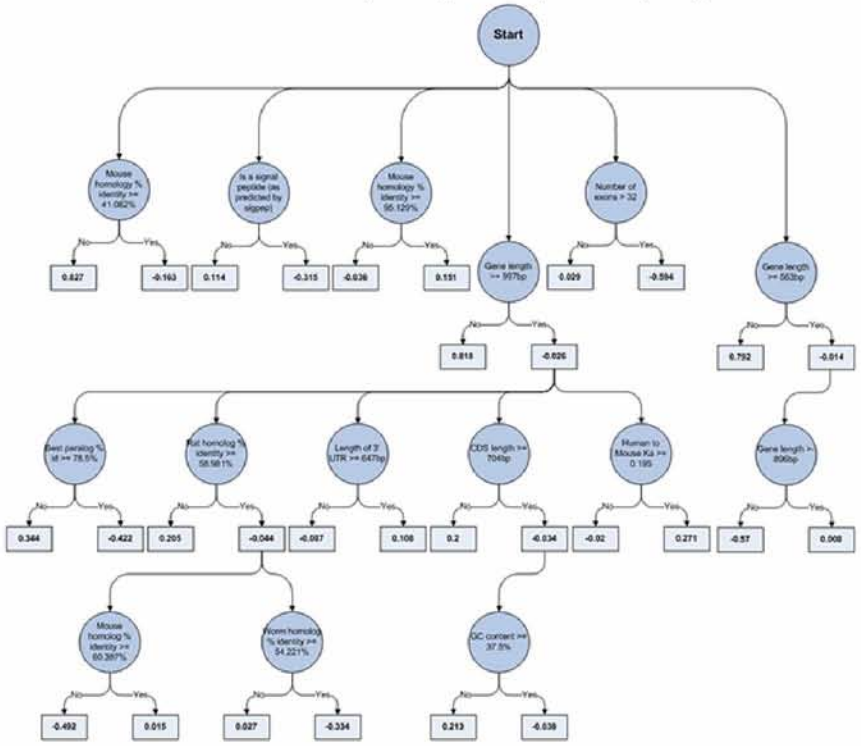
Οι αυτόματοι ταξινομητές δημιουργούνται με το να εκπαιδεύονται σε ένα σύνολο γονιδίων που έχει ήδη ταξινομηθεί με το χέρι. Τα σύνολα των υπό-εκπαιδευση γονιδίων του αλγόριθμου Prospect έχουν παραχθεί από τα 1.084 γονίδια που βρίσκονται καταχωρημένα στις βάσης δεδομένων OMIM [24] και Ensembl [44] («γονίδια που σχετίζονται με ασθένειες») και τα 1.084 γονίδια από την βάση δεδομένων Ensembl που δεν είναι γνωστό ότι εμπλέκονται σε ασθένειες («γονίδια ελέγχου») και επιλέχθηκαν τυχαία από το μεγαλύτερο σύνολο των 18.000 γονιδίων ως ένα αντιπροσωπευτικό δείγμα.

1.1.3.2 Επιλογή αλγορίθμου.

Ο αλγόριθμος Prospect χρησιμοποιεί τον αλγόριθμο Weka [46], ως την βάση για τις μηχανές πειραματικής μάθησης . Μετά από μία μεγάλη αναζήτηση για διάφορες μεθόδους μηχανικής μάθησης από τους δημιουργούς του αλγόριθμου, αποφάσισαν μαζί με τον αλγόριθμο Weka να συνδυάζεται η ανάλυση του ταξινομητή με μεθόδους δέντρου απόφασης, αφού παρέχει υψηλή ακρίβεια ζευγαριών σε σχετικά μικρό σύνολο κανόνων και χαρακτηριστικών δεδομένων [47]. Το πλεονέκτημα του δέντρου απόφασης είναι ότι βασίζεται σε σχεδιαγράμματα σε σχέση με άλλους δημοφιλείς αλγόριθμους όπως τον k-nearest [48], τις διανυσματικές μηχανές υποστήριξης και των Μπεϊσιανών δικτύων[49] είναι ότι οι κανόνες που παράγονται για τις περιπτώσεις ταξινόμησης μπορεί να ερμηνευτούν πιο εύκολα από ένα μη-ειδικό χρήστη. Αυτό ισχύει για τα δένδρα απόφασης γιατί συνήθως παράγουν δέντρα που είναι τόσο έξυπνα σχεδιασμένα και ευκολότερο κατανοητά όσο αυτά που δημιουργούνται από τους παραδοσιακούς αλγόριθμους απόφασης. Τα δέντρα απόφασης (Εικόνα 1.1.3.2) , επίσης έχουν την δυνατότητα της μέτρησης του κατά πόσο επηρεάζει ένα χαρακτηριστικό στην τελική κατάταξη ενός γονιδίου. Αυτό θα μπορούσε να αποτελέσει ένδειξη για τις ουσιώδεις διαφορές μεταξύ των γονιδίων και αυτών που είναι λιγότερο πιθανό να συμμετέχουν στη νόσο.

Τα δέντρα απόφασης δημιουργούνται με την προσθήκη κάποιων κανόνων στο δέντρο με επαναληπτικό τρόπο, έτσι ώστε να έχουν προφητική δύναμη, με την χρήση αποτελεσματικών κανόνων που προστίθενται για πρώτη φορά. Οι κανόνες αυτοί προκύπτουν αυτομάτως από τις διαφορές μεταξύ των ασθενειών και των γονιδίων ελέγχου στα εκπαιδευμένα σύνολα που παρέχονται. Ένας νέος κανόνας προστίθεται στο δέντρο, είτε ως ένας νέος "κόμβος" ή ως παιδί ενός υπάρχοντος κόμβου. Με συνδυασμό του αλγόριθμου Weka [46] , ο αριθμός των κόμβων που θα προστεθούν στο δέντρο καθορίζονται από το χρήστη πριν από την έναρξη της εκπαίδευσης. Οι πολύ λίγοι κόμβοι στο δέντρο θα είναι αραιοί, χωρίς μία αρκετά καλή συσωρευτική μεροληπτική εξουσία για να κάνει σίγουρες ταξινομήσεις. Οι πάρα πολλοί κόμβοι, αντίθετα θα οδηγήσουν σε ένα υπερβολικά περίπλοκο δέντρο όπου μεταγενέστερα οι κόμβοι με αδύναμη την μεροληπτική ενέργεια μπορεί να στρεβλώσουν το αποτελέσματα των προηγούμενων, πιο σωστών κόμβων.

Εικόνα 1.1.3.2: Παράδειγμα δένδρου απόφασης.



Στον αλγόριθμο Prospect περιορίστηκε το μέγεθος του δέντρου απόφασης σε δεκαπέντε κόμβους, όπου παρέχει μια καλή ισορροπία στην προβλεπτική του ικανότητα και στην πολυπλοκότητα. Ο κάθε κόμβος αντιπροσωπεύει ένα κανόνα, αυτό σήμαινε ότι στον αλγόριθμο Prospect θα δημιουργείται το δέντρο απόφασης με το πολύ δεκαπέντε

χαρακτηριστικά δεδομένα. Σε κάθε ένα από τα χαρακτηριστικά δεδομένα δίνεται μία τιμή, όπου στην συνέχεια μας οδηγεί στο τελικό σκορ που αντικατοπτρίζει το πόσο αξιόπιστη είναι η ιεράρχηση.

Η ιεράρχηση βασίζεται στο αποτέλεσμα του «σκορ» τελικής βαθμολογίας που συγκεντρώνει το κάθε γονίδιο. Εάν το γονίδιο έχει αρνητική τιμή είναι γενικά πιο πιθανό να σχετίζεται με μία κληρονομική ασθένεια, ενώ αν είναι το γονίδιο συγκεντρώσει μία θετική τιμή είναι γενικότερα λιγότερο πιθανό να σχετίζεται με μία κληρονομική ασθένεια.

Μέσα από της διάφορες δοκιμές που είχε ο ταξινομητής του αλγόριθμου αυτού, ήταν να δούμε κατά πόσο ανταποκρίνεται σωστά στις προβλέψεις του σε σχεδόν ανύπαρκτα δεδομένα. Έτσι πραγματοποιήθηκε ο έλεγχος «Cross-validation» (πίνακας 1.1.3.3). Αυτή η τεχνική χρησιμοποιείται ευρέως στη μηχανική μάθηση και περιλαμβάνει χωρισμό σε διαμερίσματα του συνόλου από τα δεδομένα σε δέκα ανεξάρτητες "πτυχές " όπου η κάθε μία έχει την ίδια ισορροπία από γονίδια που σχετίζονται με ασθένειες και γονίδια ελέγχου. Ο ταξινομητής εκπαιδεύεται σε εννέα από τα διαμερίσματα και δοκιμάζεται στα υπόλοιπα διαμερίσματα. Αυτό επαναλαμβάνεται μέχρι το κάθε διαμέρισμα να έχει δοκιμαστεί από ένα νέο ταξινομητή που δημιουργείται με βάση τον προηγούμενο και τα εκπαιδευμένα του σύνολα. Έτσι και σε σχεδόν ανύπαρκτα δεδομένα ο ταξινομητής μας δεν αποκλίνει από την πρόβλεψη της σωστής ανίχνευσης. Κατά μέσο όρο, το 70% των γονιδίων που σχετίζονται με ασθένειες ήταν σωστά και προσδιορίστηκαν με την τεχνική της «Cross-validation».

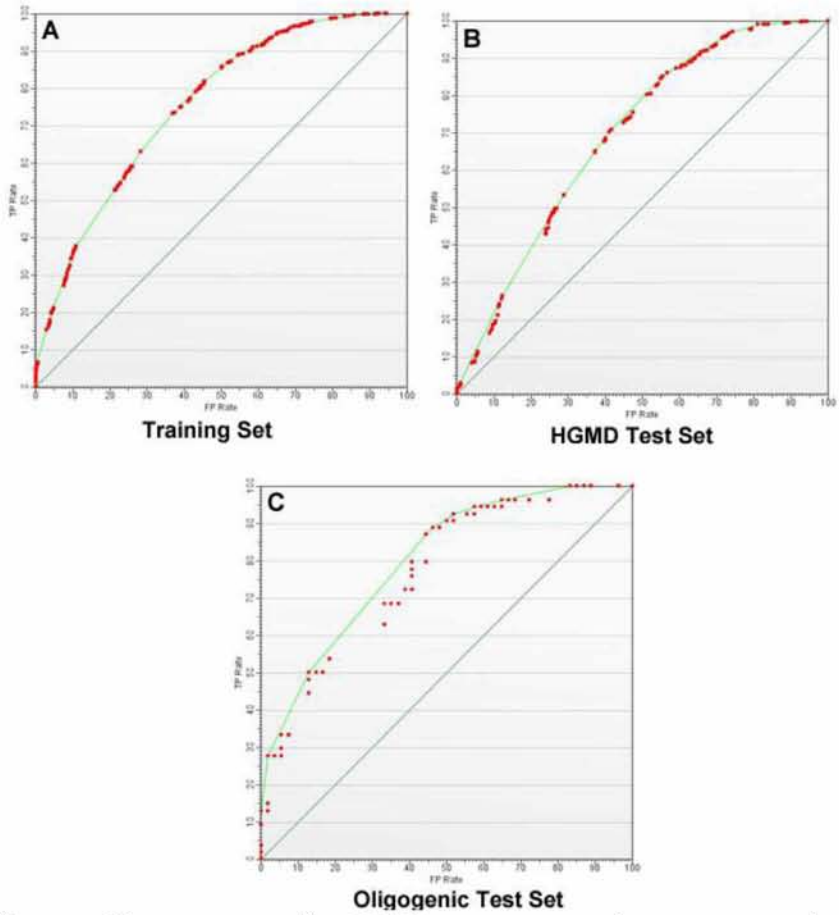
Πίνακας 1.1.3.3: Λεπτομερή στατιστικά στοιχεία σχετικά με τις επιδόσεις του ταξινομητής.

Σύνολα	Κόμβοι στο δένδρο	Ακρίβεια (Accuracy)	Ειδικότητα (Precision)	Ευαισθησία (Recall)	AUC	Kappa
Εκπαιδευμένα σύνολα (OMIM)	15	67%	65%	77%	0.75	0.35
Σύνολα «Cross validation»	15	63%	62%	70%	0.70	0.27
Σύνολα HGMD	15	64.5%	63%	71%	0.69	0.29
Ολιγογενετικά Σύνολα	15	65%	63%	72%	0.76	0.31

Όταν το δέντρο απόφασης βγάζει ένα αποτέλεσμα (σκορ) που μπορεί να είναι πολύ κοντά στο μηδέν, τότε υπάρχει ένα σχετικά απλό θέμα όσο αφορά την αύξηση της ειδικότητας σε βάρος της ευαισθησίας. Οι Receiver Operating Characteristic (ROC)

καμπύλες μπορεί να χρησιμοποιούνται για την οπτικοποίηση των επιδόσεων του ταξινομητή με διαφορετικούς συνδυασμούς της ειδικότητας και της ευαισθησίας. (Εικόνα 1.1.3.3)

Εικόνα 1.1.3.3: Receiver Operating Characteristic (ROC) καμπύλες για τα σύνολα που εξετάζουμε στον αλγόριθμο μας (εκπαιδευμένα σύνολα Α, σύνολα με γονίδια από HGMD Β, σύνολα με ολιγογενετικά γονίδια ελέγχου Γ) .



Ο άξονας Χ στην καμπύλη ROC αντιπροσωπεύει το ποσοστό των ψευδώς θετικών γονιδίων ενώ άξονας Ψ το κλάσμα των πραγματικά θετικά γονιδίων. Καθώς ο αριθμός των θετικά πραγματικών αυξάνεται (ευαισθησία), το ίδιο ισχύει για τον αριθμό των ψευδώς θετικών (μείωση ειδικότητας).

1.1.3.3 Μηχανική ταξινομητή αλγόριθμου Prospectr.

Για τον λόγο ότι μέχρι σήμερα δεν έχουμε ολοκληρωμένες μελέτες και έρευνα για όλα τα γονίδια, δεν έχουμε πλήρη εικόνα και ανάλυση μεταξύ των διαφορών που παρουσιάζουν οι ακολουθίες των φυσιολογικών γονιδίων έναντι των ακολουθιών των γονιδίων που σχετίζονται με ασθένειες. Ο αλγόριθμος Prospectr με βάση τα υφιστάμενα

δεδομένα από έρευνες που έχουν διεκπεραιωθεί μέχρι σήμερα, συγκεντρώνει όλες τις γνωστές διαφορές των ακολουθιών και σε ορισμένες περιπτώσεις εισάγει και νέες. Πέρα από τις προσπάθειες του αλγόριθμου Prospectr χρειάζεται περισσότερη έρευνα ούτως ώστε να μπορέσουμε να συνδέσουμε μεταξύ τους όλες αυτές τις διαφορές.

Το μήκος του με φορά 3 UTR πιστεύεται ότι σχετίζεται με μεταγραφική αποδοτικότητα και σταθερότητα του mRNA[50], η οποία με την σειρά της επηρεάζει το επίπεδο της έκφρασης των γονιδίων. Δύο άλλα χαρακτηριστικά σε επίπεδο ακολουθίας βρέθηκε να έχουν σημαντικές διαφορές μεταξύ των γονιδίων που σχετίζονται με ασθένειες και των γονιδίων που δεν σχετίζονται με ασθένειες. Αυτά τα χαρακτηριστικά είναι η απόσταση που έχει ένα γονίδιο με το πλησιέστερο γειτονικό γονίδιο του και ο αριθμός των εξονίων που έχει ένα γονίδιο. Επίσης θα μπορούσαν να σχετίζονται άμεσα με τα επίπεδα έκφρασης. Με αυτό τον αλγόριθμο επιβεβαιώθηκαν οι προτάσεις από τις προαναφερθέντες μελέτες ότι υπάρχει σημαντική διαφορά στην ιδιαιτερότητα των ιστών μεταξύ των γονιδίων που σχετίζονται με ασθένειες και αυτών που δεν σχετίζονται με ασθένειες.[45]

Τα γονίδια που σχετίζονται με ασθένειες πιθανός μοιράζονται χαρακτηριστικά ακολουθίας μέχρι κάποιο βαθμό. Συγκεκριμένα το μήκος του γονιδίου και το μήκος της πρωτεΐνης όταν συλλέγονται μαζί σαν ένα χαρακτηριστικό για ένα ταξινομητή δέντρου απόφασης με δεκαπέντε κόμβους μπορεί να προβλέπει με μεγάλο ποσοστό επιτυχίας (69% των γονιδίων που σχετίζονται με ασθένειες προβλέπονται σωστά και 51% παρουσιάζουν εσφαλμένη ταξινόμηση).

Μια εναλλακτική υπόθεση είναι ότι ο αλγόριθμος Prospectr δεν προβλέπει γονίδια που μπορούν να σχετίζονται με μία ασθένεια πάντα αλλά και το αντίθετο. Αντλεί την προγνωστική του ισχύει διαγράφοντας εκείνα τα γονίδια που είναι πιθανόν να μην εμπλέκονται σε ασθένειες όπως μεταλλάξεις συνήθως ως αποτέλεσμα του φαινότυπο τους που είναι είτε θανατηφόρος, είτε ανιχνεύσιμος ή πολύ εξασθενημένος.

Ο αλγόριθμος Prospectr έχει καλή απόδοση στην εξέταση ολιγογενετικών συνόλων. Ωστόσο, θα μπορούσε κανείς να περιμένει τους αιτιατούς βιολογικούς μηχανισμούς να διαφέρουν μεταξύ των απλών μεντελικών. Επίσης θα μπορούσε να διαφέρουν τα πιο πολύπλοκα χαρακτηριστικά και οι ταξινομητές που ασχολούνται με τα είδη. Μέχρι σήμερα δεν έχει εφαρμοστεί με μεγάλη επιτυχία η αναζήτηση των υποψηφίων γονιδίων που σχετίζονται με ασθένειες πολυπαραγοντικής αιτιολογίας

ασθένειες. Δεν μπορούμε να είμαστε σίγουρη κατά πόσο θα έχουμε αξιόπιστα αποτελέσματα όταν εκτελεστεί ο αλγόριθμος Prospectr για να προβλέψει διάφορα υποψήφια γονίδια που σχετίζονται με μία ασθένεια που εκδηλώνεται με τον συνδυασμό πολλών γονιδίων μαζί, δηλαδή πιο σύνθετων ασθενειών που συνδέονται με πολλά γονίδια.

1.1.3.4 Μεθοδολογία αλγόριθμου Prospectr.

Στον αλγόριθμο Prospectr [40] γίνονται οι διάφορες αναζητήσεις των υποψήφιων γονιδίων από τις βάσεις δεδομένων OMIM [24] και HGMD [51]. Από αυτές τις βάσεις δεδομένων αντλούνται οι διάφορες πληροφορίες και δημιουργούνται οι διάφοροι κατάλογοι με τα υποψήφια γονίδια με τα οποία θα ασχοληθεί ο αλγόριθμος. Για περεταίρω ανάκτηση πληροφοριών και εμπλουτισμό των γονιδίων με δεδομένα όπως πληροφορίες για την ακολουθία που έχει το κάθε γονίδιο, ο αλγόριθμος αντλεί δεδομένα από την βάση δεδομένων Ensembl [44]. Επίσης γίνεται οι χρήση της εν λόγο βάσης δεδομένων αντλώντας τυχαία γονίδια για την δημιουργία αντιπροσωπευτικών συνόλων με ευπαθή γονίδια αλλά και διάφορα άλλα σύνολα ελέγχου με γονίδια που δεν σχετίζονται με ασθένειες. Σημαντικό κομμάτι της δημιουργίας των συνόλων ελέγχου είναι τα γονίδια που τα αποτελούν. Εισάγονται με διαφορετικούς ελέγχους, άρα τα γονίδια αυτά δεν προέρχονται από τις βάσεις δεδομένων OMIM και HGMD αλλά από την βάση δεδομένων Ensembl .

Για την δημιουργία του πρωταρχικού συνόλου με τα χαρακτηριστικά ο αλγόριθμος κάνει διάφορες αναζητήσεις και συλλέγει συγκεντρωτικά διάφορες πληροφορίες από τις βάσεις δεδομένων Ensembl, HomoloGene, NCBI, Interpro[52], Swiss -PROT[25] και NGEA[53]. Μέσα σε αυτό το καινούργιο σύνολο δεδομένων με χαρακτηριστικά περιλαμβάνονται όλα τα χαρακτηριστικά που αναφέρονται στον πίνακα 1.1.3.1 καθώς και διάφορες άλλες πληροφορίες σχετικά με την έκφραση των ιστών και των πρωτεϊνικών περιοχών που καταρτίζουν την κάθε πρωτεΐνη.

Ενδεικτικά συγκρίνεται ένα χαρακτηριστικό σύνολο από 1.084 που προέρχεται από την OMIM βάση δεδομένων, με ένα αντιπροσωπευτικό δείγμα ελέγχου που απαρτίζεται από με 18.000 γονίδια που προέρχεται από την Ensembl βάση δεδομένων. Με βάση τα διάφορα χαρακτηριστικά δεδομένα που κρίθηκαν να έχουν εύλογο βαθμό προβλεπτικής ικανότητας δημιουργείται το δέντρο απόφασης. Ως βάση του αλγόριθμου Prospectr χρησιμοποιείται ο αλγόριθμος WEKA[46] και λειτουργεί ως πλατφόρμα

οικοδόμησης του ταξινομητή. Με την εκτελέσει του αλγόριθμου Weka με τα δέντρα απόφασης έχουμε ένα συνδυασμό και των δύο μεθόδων με αποτέλεσμα της δημιουργίας του Weka ADTree ταξινομητή. Με την χρήση του Weka ADTree ταξινομητή παίρνουμε εύκολα τιμές για διάφορες παραμέτρους όπως της ακρίβειας(accuracy), ειδικότητας (precision), ευαισθησίας (recall), AUG,και του στατιστικού K (Kappa).

1.1.3.5 Σχετική απόδοση αλγόριθμου Prospectr.

Ο αλγόριθμος Prospectr έχει αρκετά πλεονεκτήματα έναντι των υφισταμένων αλγόριθμων ταξινόμησης που υπάρχουν και σχεδιάστηκε να προβλέπει γονίδια που είναι περισσότερο ή λιγότερο πιθανό να σχετίζονται με μία ανθρώπινη ασθένεια. Με βάση τις πειραματικές δοκιμές που εκτελέστηκαν στον αλγόριθμο αυτό φαίνεται να εκτελείται σημαντικά καλύτερα σε αθέατα δεδομένα από το ταξινομητή που λειτουργεί με δέντρα απόφασης.

Μια από τις ιδιότητες του αλγόριθμου Prospectr, η επέκταση των δεδομένων που σχετίζονται με τα χαρακτηριστικά ακολουθίας και δομής, η διερεύνηση του περιεχομένου και της φυλογενετικής έκτασης των υποψηφίων γονιδίων επιτρέπει στους ερευνητές να κοιτάξουν τα ακριβή χαρακτηριστικά που συμβάλλουν για την συγκεκριμένη ιεράρχηση του ταξινομητή. Επιπλέον, δεν απαιτεί περαιτέρω λεπτομερή φαινοτυπική γνώση για την εν λόγω ασθένεια και μπορεί να σκοράρει ολόκληρα τα χρωμοσώματα σε λίγα μόνο λεπτά. Η χρήση των χαρακτηριστικών ακολουθίας αποφεύγει τις μεροληπτικές συνδέσεις με τρέχων λειτουργικούς σχολιασμούς , όπου τα καλύτερα εκπαιδευμένα γονίδια είναι πολύ πιο πιθανό να έχουν καλύτερο και πιο εκτενή σχολιασμό.

Ο αλγόριθμος Prospectr χρησιμοποιήθηκε για να σκοράρει κάθε γνωστό γονίδιο της Ensembl βάσης δεδομένων σχετικά με την πιθανότητα ότι θα σχετίζεται με κάποια ανθρώπινη κληρονομική νόσο. Ομαλοποίησαν το σκορ a που έχει δοθεί από τον ταξινομητή σε κάθε γονίδιο με την εξίσωση:

$$1/1 + \gamma^{\alpha}$$

όπου γάμμα (γ) αντιπροσωπεύει την σταθερά Euler, έτσι ώστε να παίρνουμε τιμές μεταξύ 0 και 1 με την υψηλότερη βαθμολογία που υποδηλώνει μια υψηλότερη πιθανότητα το γονίδιο να σχετίζεται με την νόσο.

1.1.3.6. Εξάλειψη και των πηγών από όπου δημιουργούν σφάλματα.

Στο αλγόριθμο Prospectr συγκρίθηκε ο αριθμός των πρωτεϊνικών περιοχών μεταξύ των γονιδίων της βάσης δεδομένων OMIM και μιας ομάδας άλλων γονιδίων που δεν είναι γνωστό ότι εμπλέκονται με μία ασθένεια και δεν διαπιστώθηκε να έχουν σημαντικές διαφορές. Επομένως ο αλγόριθμος αυτός καταρρίπτει τέτοιου είδους χαρακτηριστικά πρωτεϊνικών περιοχών γιατί πιστεύει ότι τα εν λόγο χαρακτηριστικά είναι μεροληπτικά και θα προκαλέσουν θόρυβο και σφάλματα στα αποτελέσματα μας.

Αν και εντοπίστηκαν εξαιρετικά σημαντικές διαφορές στην έκφραση των διαφόρων προτύπων ιστών που εντοπίστηκαν στον αλγόριθμο Prospectr , αποκλείστηκε και αυτό το χαρακτηριστικό από την πρόγνωση και ιεράρχηση διαφόρων γονιδίων που σχετίζονται με ασθένειες. Ο λόγος για τον οποίο δεν συμπεριλαμβάνεται στα χαρακτηριστικά δεδομένα του αλγόριθμου τα πρότυπα των ιστών είναι γιατί παρά το ότι με βάση έρευνες υπήρχαν διαθέσιμα δεδομένα έκφρασης περίπου 95% των γονιδίων που σχετίζονται με ασθένειες αλλά για δεδομένα ελέγχου υπήρχαν δεδομένα μόλις 60% για τα γονίδια ελέγχου, και για αυτό το είδος χαρακτηριστικών δεδομένων χαρακτηρίστηκαν μεροληπτικά και δεν συμπεριλήφθηκαν στο σύνολο των χαρακτηριστικών δεδομένων.

Ακόμη ένα χαρακτηριστικό που πιθανός προκαλεί μεροληψία είναι ο βαθμός φυλογενετικής ανάλυσης. Ο βαθμός φυλογενετικής ανάλυσης πηγάζει από την εξέταση τον ομολόγων, όμως για τον λόγο ότι στον συγκεκριμένο τομέα υπάρχουν περισσότερο από το επιτρεπτό όριο δεδομένα, καταλαβαίνουμε ότι δεν είναι πολύ καλά σχολιασμένα με αποτέλεσμα να μην έχουμε μια καθαρή πρόβλεψη για το γονίδιο που εξετάζουμε με συνέπεια όχι ακριβή πρόβλεψη οντολογιών [54]. Στον αλγόριθμο Prospectr μέχρι και το 34% των προβλεπτικών ικανοτήτων προέρχεται από τα χαρακτηριστικά που συνδέονται με το βαθμό φυλογενετικής ανάλυσης.

1.1.4 Αλγόριθμος Suspects.

Στόχος του αλγόριθμου Suspects είναι η αποτελεσματική αυτοματοποίηση των πρώτων βημάτων [55] στην προσέγγιση υποψήφιων γονιδίων που εμπλέκονται στην εκδήλωση μίας ασθένειας. Μερικά από τα πρώτα βήματα που αυτοματοποιεί ο αλγόριθμος αυτός είναι:

-Προσδιορισμός των γενετικών παραγόντων των ανθρώπινων φαινότυπων. Ορισμένοι μεντελικοί φαινότυποι μπορεί να θεωρηθούν ως κανονικές παραλλαγές και όχι ασθένειες. Όμως πολλές γενετικές παραλλαγές συμβάλλουν κατά πολύ στην εκδήλωση διαφόρων ασθενειών και για αυτό, σκοπός του αλγόριθμου αυτού είναι να προσδιορίσει όλα τα γονίδια που σχετίζονται με ασθένειες με βάση τον φαινότυπο τους, αλλά και να ανακαλύψει όλες τις γενετικά φαινοτυπικές παραλλαγές που παρουσιάζουν, όπως παραλλαγές στην αλληλουχία των αζωτούχων βάσεων του DNA. Με τον προσδιορισμό αυτών των γενετικών παραγόντων των ανθρώπινων φαινοτύπων ο αλγόριθμος θα μπορεί να συνδυάσει τα δεδομένα και να ταιριάξει όλα τα γονίδια που σχετίζονται με ασθένειες που παρουσιάζουν τις ίδιες γενετικές παραλλαγές.

- Προσδιορισμός γονιδίων που σχετίζονται με ασθένειες με βάση τα αποτελέσματα της μεθόδου «positional cloning». Είναι μία μέθοδος ταυτοποίησης του γονιδίου για ένα συγκεκριμένο φαινότυπο που ταυτίζεται σε μία συγκεκριμένη χρωμοσωμική θέση. Η υποψήφια χρωμοσωμική περιοχή απομονώνεται με την χρήση μεθόδων κλωνοποίησης μέχρι τον εντοπισμό του γονιδίου και των μεταλλάξεων του.

-Εντοπισμός γονιδίων που σχετίζονται με ασθένειες με βάση την ανεξάρτητη θέση τους. Η μέθοδος αυτή βασίζεται στην ομολογία της ακολουθίας αλλά και στην λειτουργικότητα της, όταν αναζητούμε σε προκαθορισμένη χρωμοσωμική περιοχή και όχι σε ολόκληρο το ανθρώπινο γονιδίωμα. Οι ομόλογες αναζητήσεις έχουν πολύ καλά αποτελέσματα όταν συνδυάζονται με πληροφορίες θέσης.

-Ταυτοποίηση ενός γονιδίου που σχετίζεται με μία ασθένεια με βάση της γνωστής του πρωτεΐνης. Εκτελείται με την χρήση συγκεκριμένων ολιγο-νουκλεοτιδίων του γονιδίου(εύρεση της πρωτεϊνικής αλληλουχίας από αμινοξέα, κατά συνέπια εύρεση της cDNA αλληλουχίας και συσχετισμός της με τις υπόλοιπες διαθέσιμες που υπάρχουν στις cDNA βιβλιοθήκες) ή ειδικών αντισωμάτων που μπορούν να χρησιμοποιηθούν για τον εντοπισμό του γονιδίου.

-Ταυτοποίηση ενός γονιδίου που σχετίζεται με ασθένειες με βάση την γνωστή του DNA αλληλουχίας.

-Εντοπισμός ενός γονιδίου που σχετίζεται με μία ασθένεια με βάση τα λειτουργικά χαρακτηριστικά του.

-Εντοπισμός υποψήφιων γονιδίων που σχετίζονται με ασθένειες από ένα συνδυασμό της γονιδιακής τους έκφρασης, λειτουργικότητας και των ομόλογων τους.

- Επιβεβαίωση του υποψηφίου γονιδίου ότι σχετίζεται με μια ασθένεια αλλά και κατανόηση της λειτουργικότητας του.

Ο αλγόριθμος Suspects σε κάθε αναζήτηση του συσχετίζει τις πληροφορίες που φέρει κάθε γονίδιο, όπως ορολογίες από την βάση δεδομένων GO[21, 30], πρωτεϊνικές περιοχές[52] και πρότυπα έκφρασης των γονιδίων που έχουν υλοποιηθεί πρώτα από τον αλγόριθμο ιεραρχικής αναζήτησης Prospectr [40] που περιγράφουμε πιο πάνω. Δηλαδή είναι μία περεταίρω επέκταση των αποτελεσμάτων που μας παράγει ο αλγόριθμος Prospect.

1.1.4.1 Μεθοδολογία αλγόριθμου Suspects .

Ο αλγόριθμος αυτός λειτουργεί μη βάση την προϋπόθεση ότι τα γονίδια που εμπλέκονται σε ένα σύνθετο χαρακτηριστικό θα ανήκουν σε παρόμοιες περιοχές και επομένως θα είναι περισσότερο πιθανό να μοιράζονται πανομοιότυπες πρωτεϊνικές περιοχές, περιγραφές-σχολιασμούς που αφορούν τα γονίδια και πρότυπα έκφρασης. Σε μία αναζήτηση έχουμε εισαγωγές από δεδομένα όπως λέξεις κλειδιά για την ασθένεια ή την γονιδιακή περιοχή που μας ενδιαφέρει. Μπορούμε να ήμαστε πιο συγκεκριμένη σε αυτήν την εισαγωγή δεδομένων εισάγοντας δείκτες ,ζώνες, χρωμοσωμικές περιοχές ή ακόμη και γονίδια.

-Με βάση τις λέξεις κλειδιά που εισάγουμε ο αλγόριθμος Prospects κάνει τις δικές του αναζητήσεις για να μαζέψει τα υποψήφια γονίδια που σχετίζονται με τις λέξεις κλειδιά από τις βάσεις δεδομένων OMIM[24], HGMD[51] και GAD[56].

-Αφού εκτελέσει τις δικές του αναζητήσεις τότε συνεργάζεται με τον αλγόριθμο Prospectr όπου και αυτός του επιστρέφει μια λίστα με τα υποψήφια γονίδια που βρίσκει με βάση τις λέξεις κλειδιά που εισάγαμε. Παίρνει τα υποψήφια γονίδια που του επιστρέφει ο αλγόριθμος Prospectr και μαζί με τα γονίδια που έχει αναζητήσει από τις πιο πάνω βάσεις δεδομένων δημιουργεί μία λίστα με όλα τα αστάθμιστα γονίδια.

Με την δημιουργία της τελικής λίστας με όλα τα υποψήφια γονίδια, τότε τα σκοράρει με βάση τις πληροφορίες των GO όρων, πρωτεϊνικών περιοχών και των προτύπων γονιδιακής έκφρασης που έχει για το κάθε γονίδιο με τα ακόλουθα βήματα:

(i)Για κάθε υποψήφιο γονίδιο που συμπεριλαμβάνεται στην δική του λίστα με τα υποψήφια γονίδια που σχετίζονται με μία ασθένεια, ο αλγόριθμος αναζητά τις GO ορολογίες που χαρακτηρίζουν το καθένα γονίδιο που είναι εννοιολογικά παρόμοιες σε σημαντικό επίπεδο. Ακολούθως εκτελεί τους μεταξύ τους συσχετισμούς και συνδέει τα γονίδια ανάλογα με το πόσο ταιριάζουν μεταξύ τους με βάση τους όρους που έχει το καθένα. Το κάθε γονίδιο βαθμολογείται με βάση το πόσο πολύ βρέθηκε να ταιριάζει με τους όρους των άλλων γονιδίων που βρίσκονται στην λίστα με τα υποψήφια γονίδια. (p-valueGO)

(ii)Μετά εκτελεί αναζητήσεις και σκοράρει κατά τον ίδιο τρόπο τα γονίδια αλλά αυτή την φορά με βάση τις πρωτεϊνικές περιοχές που περιέχονται στο κάθε γονίδιο και εκτελεί πάλι τους μεταξύ τους συσχετισμούς για να παράγει ένα δεύτερο τρόπο βαθμολόγησης του κάθε γονιδίου. (p-valueInterPo)

(iii)Παρόμοια με τους δύο πιο πάνω τρόπους σκοραρίσματος υπολογίζει την βαθμολογία συσχετισμού του κάθε γονιδίου με τα υπόλοιπα αλλά αυτή την φορά με βάση το πρότυπο γονιδιακής έκφρασης που έχει το κάθε γονίδιο αλλά και το κατά πόσο καλά συνδέονται μεταξύ τους. (p-valueExpression)

Αφού εκτελέσει όλους τους συσχετισμούς μεταξύ των γονιδίων και υπολογίσει και τις τρείς βαθμολογίες για το κάθε γονίδιο, τότε υπολογίζει την τελική βαθμολογία (Total_p-value) που συγκεντρώνει το κάθε γονίδιο με τον ακόλουθο τύπο:

$$Total\_p\text{-}value = (p\text{-}valueGO + p\text{-}valueInterPo + p\text{-}valueExpression) / 3$$

Ακολούθως ιεραρχεί τα υποψήφια γονίδια με βάση την τελική βαθμολογία που συγκέντρωσε το καθένα και εξάγει την τελική λίστα με όλα τα πιθανά ταξινομημένα γονίδια που σχετίζονται με την ασθένεια που αναζητεί.

1.1.5 Αλγόριθμος SNPs3D .

Η σχέση μεταξύ των ευπαθείς ασθενειών και των γενετικών παραλλαγών είναι πολύπλοκη, αλλά για πολλούς διαφορετικούς τύπους δεδομένων είναι συναφή. Ο αλγόριθμος SNPs3D[57] είναι μία διαδικτυακή πηγή αναζήτησης σε βάσεις δεδομένων η οποία παρέχει και ενσωματώνει όσο το περισσότερο δυνατών πληροφορίες για ασθένειες που σχετίζονται με γονίδια σε μοριακό επίπεδο.

Για καλύτερη κατανόηση του μπορούμε να τον χωρίσουμε σε τρεις κύριες ενότητες. Μία ενότητα είναι να προσδιορίζει ποια γονίδια είναι υποψήφια στο να σχετίζονται σε μία ασθένεια. Μια δεύτερη ενότητα είναι να παρέχει σχετικές πληροφορίες για τις σχέσεις ανάμεσα στα σύνολα των υποψήφιων γονιδίων. Η τρίτη ενότητα αναλύει τις πιθανές περιπτώσεις μη-συνώνυμων SNPs σε επίπεδο λειτουργίας πρωτεϊνών. Τα υποψήφια γονίδια που σχετίζονται με ασθένειες και οι σχέσεις μεταξύ γονιδίων προκύπτουν από μελέτες με απλές αλλά αποτελεσματικές δημοσιεύσεις. Οι σχέσεις των SNP με τις λειτουργίες των πρωτεϊνών παράγονται από δύο μεθόδους, η πρώτη μέθοδος με βάση την δομής και σταθερότητα της πρωτεΐνης, και η άλλη μέθοδος με βάση τις καλά διατηρημένες ακολουθίες των πρωτεϊνών. Οι ενδείξεις για κάθε γονίδιο περιλαμβάνουν μια σειρά από συνδέσεις με άλλα δεδομένα, όπως τα πρότυπα έκφρασης, τα περιεχόμενα των πλαισίων(path ways), πειραματικά δεδομένα και δημοσιεύσεις. Οι αλληλεπιδράσεις των γονιδίων παρουσιάζονται σε με μία διαδραστική διασύνδεση γραφημάτων, για γρήγορη πρόσβαση στις βασικές πληροφορίες.

1.1.5.1 Γενική περιγραφή λειτουργίας του αλγόριθμου SNPs3D..

Η ανάλυσης των SNP, βασίζεται στο γνωστό δίκτυο του κάθε γονιδίου, στις σχέσεις μεταξύ των υποψήφιων γονιδίων, καθώς και στην πρόσβαση σε ένα ευρύ φάσμα από δεδομένα που προέρχονται από τις δημοσιεύσεις. Έτσι ο χρήστης έχει την δυνατότητα να αφομοιώσει γρήγορα τις διαθέσιμες πληροφορίες, και να ανάπτυξη διάφορα μοντέλα από τις αλληλεπιδράσεις των γονιδίων που σχετίζονται με ασθένειες.

Η σύνδεση των μελετών, στηρίζεται στην ανάλυση των γενετικών διαφορών, ιδίως μεταξύ των SNPs που σχετίζονται ή δεν σχετίζονται με μια ασθένεια σε ένα ευρύτερο πληθυσμό. Είναι περισσότερό ισχυρή η ανίχνευση των εν λόγω χαμηλών σημάτων. Περίπου 10 εκατομμύρια ανθρώπινα SNPs έχουν μέχρι στιγμής εντοπιστεί [58]. Επί του παρόντος, οι μελέτες σύνδεσης εξαρτώνται από την επιλογή ενός

υποσυνόλου από αυτές και επηρεάζουν την πιθανότητα τα υποψήφια γονίδια να σχετίζονται με την ασθένεια ή βρίσκονται σε ανισορροπία με την μεταξύ τους σύνδεση.

Πρωταρχικός σκοπός του αλγόριθμου SNPs3D είναι να παρέχει ένα μέσο για την επιλογή των υποψήφιων γονιδίων που πιθανός να επηρεάζουν την ευπάθεια της νόσου, και να επιλέξει τα περαιτέρω σημαντικότερα μη συνώνυμα SNPs στο πλαίσιο αυτών των γονιδίων. Η ταχεία συσσώρευση νέων δεδομένων σχετικά με τα ανθρώπινα SNPs, η ολοκλήρωση της ακολουθίας ολόκληρου του ανθρώπινου γονιδιώματος, και η αύξηση των πληροφοριών σχετικά με τις βιομετρικές αλληλεπιδράσεις οδηγεί σε ένα καλύτερο και αξιόπιστο μηχανισμό κατανόησης της σχέσης μεταξύ γονότυπου και νόσου. Επί του παρόντος, οι σχετικές πληροφορίες είναι ακόμα πολύ ελλιπείς, και είναι διάσπαρτες σε όλες τις βάσεις δεδομένων και σε χιλιάδες άρθρα.

Κατά δεύτερο σκοπό ο αλγόριθμος αυτός συλλέγει και να ενσωματώσει όσο το δυνατόν περισσότερα δεδομένα σε μοριακό επίπεδο σχετικά με τους μηχανισμούς σύνδεσης και γενετικής τροποποίησης της ασθένειας. Για την επίτευξη αυτών των στόχων, ο αλγόριθμος είναι οργανωμένος σε τρεις ενότητες. Η πρώτη οντότητα παράγει τους καταλόγους των υποψηφίων γονιδίων για οποιαδήποτε συγκεκριμένη ασθένεια, κάνοντας χρήση την ανάλυση της σχέσης μεταξύ της νόσου και των γονιδίων, όπως αντικατοπτρίζονται στις δημοσιεύσεις. Η δεύτερη ενότητα παρέχει ένα γραφικό δίκτυο των αλληλεπιδράσεων μεταξύ των γονιδίων, που δημιουργείται με τις συσχετίσεις μεταξύ των δημοσιεύσεων, τις γνωστές αλληλεπιδράσεις μεταξύ των πρωτεϊνών[12, 59], και τις υπάρχουσες οδούς (path ways)[60, 61]. Η τρίτη ενότητα παρέχει πληροφορίες για τη σχέση μεταξύ των μη συνώνυμων SNPs και της λειτουργίας των πρωτεϊνών.

Ο εντοπισμός των υποψήφιων γονιδίων και η κατασκευή των αντιπροσωπευτικών δικτύων του κάθε γονίδιου γίνονται με την χρήση απλών τεχνικών εξόρυξης κειμένου. Κατασκευάζονται προφίλ για κάθε ασθένεια και για κάθε γονίδιο. Κάθε προφίλ (ασθένεια ή γονίδιο) εκπροσωπείται από μια λίστα με λέξεις κλειδιά και συνδέεται με τις πιο στενές τους έννοιες. Το σύνολο των λέξεων και των όρων παράγεται από το περιεχόμενο των 80.000 PubMed[39] αποσπασμάτων που συνδέονται με ένα ή περισσότερα ανθρώπινα γονίδια με βάση την NCBI[24] βάση δεδομένων. Τα ζεύγη εννοιών, όπως δύο γονίδια ή μια ασθένεια και ένα γονίδιο, συνδέονται με την επικάλυψη των λέξεων κλειδιών που περιέχουν στα προφίλ τους. Έτσι δημιουργείται ένα δίκτυο που προκύπτει από την μεταξύ σχέση των γονιδίων, δεδομένου ότι απορρέει άμεσα τις

πληροφορίες που προέρχονται από τις δημοσιεύσεις. Μια ποικιλία από άλλες υπολογιστικές μεθόδους αναπτύσσονται για την εξαγωγή αυτόματα των πληροφοριών από τις δημοσιεύσεις. Μέχρι σήμερα, αυτές οι μέθοδοι δεν έχουν χρησιμοποιηθεί εκτενώς στις εν γένει διαθέσιμες διασυνδέσεις. Ο αριθμός των ομάδων, αναφέρεται στην Ingenuity Pathway βάσης δεδομένων και στην Protein Reference Βάση δεδομένων [62], Αν και αυτές οι βάσεις δεδομένων παρέχουν ακριβή δεδομένα, ο ανθρώπινος παράγοντας για την επεξεργασία και επιμέλεια, καθιστά την όλη διαδικασία και ανάπτυξη αργή. Αυτό το πρόβλημα γίνεται ακόμη πιο σοβαρό, με τον ρυθμό με τον οποίο αυξάνονται οι δημοσιεύσεις. Οι αλληλεπιδράσεις μεταξύ πρωτεΐνης και δικτύων κατασκευάζεται επίσης αυτόματα[63-66], χρησιμοποιώντας μοντέλα πιθανοτήτων ενσωματώνοντας πειραματικά δεδομένα, όπως οι μέθοδοι yeast-2-hybrid [67, 68] και TAP pull-downs[69].

Στον αλγόριθμο SNPs3D, οι πιθανές λειτουργικές επιπτώσεις των μη συνώνυμων SNPs εκτιμούνται με τη χρήση δύο αναπτυγμένων μεθόδων[70-72]. Η μια μέθοδος κάνει χρήση της δομής της πρωτεΐνης για να εξακριβώσει ποιές αντικαταστάσεις αμινοξέων σημαντικά αποσταθεροποιούν την δομή της. Τα αποτελέσματα δείχνουν ότι έως και τα τρία τέταρτα των μονογονιδιακών νόσων /μεταλλάξεων ενεργούν κατ 'αυτόν τον τρόπο[71]. Η δεύτερη μέθοδος προσδιορίζει τις επιβλαβής αντικαταστάσεις μέσω της ανάλυσης, της διατήρησης της θέσης των αμινοξέων που επηρεάζουν την ακολουθία[72].

Επίσης έχουν αναπτυχθεί μέθοδοι για την αξιολόγηση του μοριακού αποτελέσματος των μη συνώνυμων SNPs[73-80]. Ορισμένες από τις μεθόδους αυτές αποτελούν τη βάση των εργαλείων και των σχετικών αναλύσεων που είναι διαθέσιμα μέσω των web servers. Χρησιμοποιούνται διάφοροι αλγόριθμοι για καθορισμό των χρήσιμων πληροφοριών από διάφορα εργαλεία ώστε ο αλγόριθμος SNPs να απεικονίσει την τρισδιάστατη οπτικοποίηση τους, όπως MutDB[81] , TopoSNP[82] , SAAP[83], με λεπτομερείς ανάλυση των επιπτώσεων των μοριακών nsSNPs.

Ο αλγόριθμος SNPs3D στοχεύει στην ενσωμάτωση όλων των διαθέσιμων σχετικών στοιχείων για την αξιολόγηση της πιθανής σχέσεις συγκεκριμένων γονιδίων και SNPs σε μια ασθένεια. Έμφαση δίνεται στο να παρέχεται στους χρήστες πρόσβαση σε όλες τις βασικές πληροφορίες, έτσι ώστε να μπορούν να προβαίνουν σε ενημερωμένες αποφάσεις. Για το σκοπό αυτό, εκτός από την ανάλυση των SNP, παρέχονται σύνδεσμοι σε σχετικές αναλύσεις όπως, η GAD[56], OMIM[24], HGMD[58], GO[84], πρότυπα έκφρασης[53] και πειραματικά δεδομένα.

1.1.5.2 Κατασκευαστικά και περιεχόμενο αλγόριθμου SNPs3D / Λίστα Διεπαφών.

Κάθε μια από τις τρεις ενότητες (SNP ανάλυση, δημιουργία δικτύου γονιδίων, εξαγωγή λιστών και δικτύων με υποψήφια και ιεραρχημένα γονίδια που σχετίζονται με τις νόσους) είναι προσβάσιμες μέσω ενός απλού παράθυρο αναζήτησης, στο χώρο της πρώτης σελίδας.

Η αναζήτηση υποψήφιου γονιδίου παίρνει οποιαδήποτε λέξη ή φράση, όπως μια εγγραφή, και καταρτίζει ένα προφίλ γύρω από την έννοια αυτή. Για την ανάλυση των SNP και των αιτημάτων που δημιουργούνται από τα γονιδιακά δίκτυα, μία ιεραρχική συμβολοσειρά (λέξη κλειδί) από ερωτήματα χρησιμοποιείται, παρέχοντας μια ευρεία επιλογή από τύπους ονομάτων ή συνώνυμων λέξεων κλειδιών, συμπεριλαμβανομένων dbSNP ID, Entrez Gene ID, RefSeq ID, NBCI Gene λέξεις, και κοινών ονομασιών πρωτεϊνών, με την ακόλουθη διαδικασία:

(i) Οι λέξεις-κλειδιά επιθεωρούνται για να διαπιστωθεί εάν η σύνθεσή τους είναι συμβατή με το dbSNP ID, Entrez Gene ID ή RefSeq ID. Αν ένα από αυτά τα είδη ονομάτων διαπιστωθεί, τότε η διαδικασία εκτελείται ξανά κατά την αντίστοιχη λίστα των συνώνυμων που βρίσκουμε, και μετά από αρκετές αναζητήσεις στις διάφορες βάσης, τα αποτελέσματα με τις κατάλληλες λέξεις κλειδιά επιστρέφονται.

(ii) Εάν ο τύπος της ταυτότητας δεν μπορεί να προσδιοριστεί, τότε κάνοντας χρήση τις λέξεις-κλειδιά εκτελεί αναζητήσεις στην βάση δεδομένων NCBI[37] για να ανακαλύψει τα γονίδια που σχετίζονται με αυτές, και έτσι να ανακτήσει από εκεί τα αντίστοιχα ID για τις υπόλοιπες βάσης δεδομένων. Όταν η ακριβής αντιστοιχία βρίσκεται, τα αποτελέσματα επιστρέφονται.

(iii) Εάν οι λέξεις-κλειδία δεν ταιριάζουν με καμία λέξη κλειδί στην NCBI βάση δεδομένων τότε γίνεται μια αναζήτηση σε όλες τις περιλήψεις των δημοσιεύσεων από την βάση δεδομένων PubMed , μαζεύει τις συνώνυμες λέξεις κλειδιά και ακολουθεί την παραπάνω διαδικασία.

(iv) Αν η αναζήτηση αποτύχει τελείως, προσφέρεται στο χρήστη μια εναλλακτική λύση στο παράθυρο αναζήτησης, με σαφείς κατηγορίες σε λέξεις κλειδιά και IDs για κάθε ασθένεια.

1.1.5.3 Κατασκευαστικά και περιεχόμενο αλγόριθμου SNPs3D / Σύνολα δεδομένων από δημοσιεύσεις.

Οι περιλήψεις όλων των εγγραφών που συνδέονται με MEDLINE για κάθε γονίδιο στη βάση δεδομένων NCBI Gene [56] είναι η πηγή για τις λέξεις κλειδιά και όρους. Στην τρέχουσα έκδοση, υπάρχουν, 80.249 Medline αναφορές που συνδέονται με 19.228 ανθρώπινα γονίδια. Ο αριθμός των εμφανίσεων της κάθε λέξεις-κλειδί «KW» σε όλες τις περιλήψεις («Total_count (KW)») κρατάτε, καθώς και ο αριθμός των εμφανίσεων της κάθε λέξεις κλειδί στις περιλήψεις που συνδέονται με κάθε γονίδιο «G», «Count (G, KW)», αλλά και το κλάσμα όλων των εμφανίσεων της κάθε λέξεις-κλειδί που σχετίζονται με κάθε γονίδιο και υπολογίζεται ως εξής :

$$F1\ (G, KW) = Count(G, KW) \,/\, Total(KW)$$

1.1.5.4 Κατασκευαστικά και περιεχόμενο αλγόριθμου SNPs3D / Πίνακας σχέσεις γονιδίων.

Η σχέση αλληλεπίδρασης *L(i,j)* μεταξύ κάθε ζεύγους γονιδίων i και j υπολογίζεται ως εξής:

$$L\ (i, j) = \sum_{KW} \quad F1(Gi, KW) + \sum_{KW} \quad F1(GJ, KW)$$

όπου *L (i, j)* το άθροισμα όλων των λέξεων κλειδιών που είναι κοινά στα δύο γονίδια i,j αποκλείοντας αυτά που βρέθηκαν σε περισσότερα από 300 γονίδια. Περισσότερο μελετήθηκαν τα γονίδια που έχουν δημοσιεύσεις και συνδέονται με την βάση δεδομένων NCBI[37]. Η σύγκριση με μια πιο ισότιμη στάθμιση γονιδίων, με βάση το άθροισμα τους χρησιμοποιείται για να σύνδεση της ασθένειας με τα γονίδια. Έμφαση πρέπει να δίνεται στο κομβικό σημείο αλλά και τα γονίδια είναι χρήσιμο να συμπεριλαμβάνονται με τις σχετικές τους συνδέσεις, αλλά και τα υψηλά συζευγμένα γονίδια.

Για γρηγορότερη επεξεργασία των δεδομένων, οι σχετικές αλληλεπιδράσεις είναι αποθηκευμένες ως ένας πίνακας, διατηρώντας το μεγαλύτερο όριο των 200 αλληλεπιδρά-σεων. Ωστόσο, σε όλες σχεδόν τις περιπτώσεις, αυτά τα στοιχεία θα πρέπει να συμπεριληφθούν στον κατάλογο συσχετίσεων για τα άλλα γονίδια.

1.1.5.5 Κατασκευαστικά και περιεχόμενο αλγόριθμου SNPs3D / Δημιουργία λίστας με υποψήφια γονίδια που σχετίζονται με μία ασθένεια.

Δίνοντας ένα όνομα ασθένειας, μια λίστα των υποψήφιων γονιδίων παράγεται ως εξής:

Α. Εντοπίζεται το υποσύνολο των περιλήψεων που σχετίζονται με την ασθένεια:

(i) Οποιαδήποτε περίληψη που περιέχει το πλήρες όνομα της ασθένειας, για παράδειγμα, «Ο καρκίνος του μαστού».

(ii) Εάν αυτή η διαδικασία έχει ως αποτέλεσμα λιγότερο από 20 περιλήψεις, και το όνομα της ασθένειας αποτελείται από περισσότερες από μία λέξη, συνδυάζονται οι λέξεις που αποτελούν την λέξη κλειδί και εκτελείτε ξανά η αναζήτηση. Δηλαδή, για παράδειγμα, «στήθος» και «καρκίνος».

(iii) Εάν τα εβρισκόμενα αποσπάσματα είναι λιγότερο από δέκα, η διαδικασία ματαιώνεται και επιστρέφει ένα μήνυμα "Δεν υπάρχουν αρκετές δημοσιεύσεις για την κατασκευή ενός προφίλ ».

Β: Προφίλ με βάση τις λέξεις κλειδιά που δημιουργούνται για την ασθένεια, χρησιμοποιώντας τα επιλεγμένα αποσπάσματα.

Όλοι οι βασικοί όροι κατατάσσονται με το κλάσμα των περιλήψεων που έχουν στις δημοσιεύσεις για κάθε ασθένεια που περιέχουν:

$$Rank\ (KW) = Count\_abstracts\ (D, KW)\ /[Total\_abstracts\ (KW)+50]$$

όπου Count_abstracts(D, KW) είναι ο αριθμός των περιλήψεων για τη νόσο του «D» που περιέχει τις KW λέξεις κλειδιά, και Total_abstracts(KW) είναι ο συνολικός αριθμός των περιλήψεων που περιέχουν τις λέξεις κλειδιά. Με την καταμέτρηση των ψευδώς αρνητικά περιλήψεων προστίθεται 50 για τη μείωση του θορύβου. Οι πρώτες 40 κορυφαίες λέξεις κλειδιά-όροι επιλέγονται παρέχοντας κατάταξη (KW) που φτάνει τουλάχιστον το 0,1.

Γ: Η επικάλυψη των λέξεων κλειδιών – βασικών ορών της ασθένειας με εκείνους του κάθε γονίδιου υπολογίζεται:

(i) Ο αριθμός των επαναλήψεων που κάθε επιλεγμένη λέξη κλειδί- όρος«KW», εμφανίζεται σε δημοσιεύσεις που σχετίζονται με τη νόσο «D», το Count (D,KW) προσδιορίζεται και η σχετική συχνότητα υπολογίζεται όπως:

$$F2\ (D, KW) = Count\ (D, KW)\ /\ Total\_Count\ (KW)$$

(ii) Το ποσοστό της συσχέτισης της ασθένειας «D» με ένα γονίδιο «G» υπολογίζεται ως το γινόμενο της σχετικής συχνότητας των βασικών όρων-λέξεων κλειδιών της ασθένειας με τη σχετική συχνότητα των ίδιων των βασικών όρων-λέξεων κλειδιών σε αυτό το γονίδιο και δίνεται από :

$$SD\ (D, G) = \sum_{KW} F1(G, KW) * F2(D, KW)$$

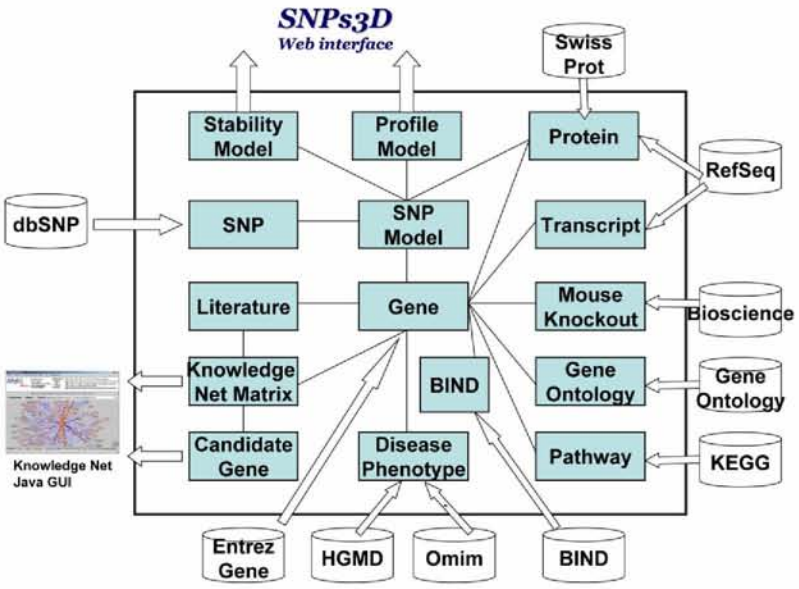
όπου παίρνουμε τους πρώτους 40 όρους-λέξεις κλειδιά όπως έγινε και στο σύνολο των όρων-λέξεων κλειδιών της νόσου «D».

Δ: Τέλος, όλα τα γονίδια με μια μη μηδενική βαθμολογία, επιστρέφονται ως υποψήφια.

1.1.5.6 Δημιουργία Βάσης δεδομένων του αλγόριθμου SNPs3D.

Η βάση δεδομένων υλοποιείται σε MySQL . Όπως φαίνεται στην Εικόνα 1.1.5.6.1, ο κεντρικός πίνακας είναι το «Γονίδιο», και ο κατάλογος με τα ανθρώπινα γονίδια δημιουργείται με δεδομένα από την NCBI[37] βάση δεδομένων. Ο πίνακας Γονίδιο συνδέεται με άλλους εξειδικευμένους πίνακες. Ο πίνακας SNP περιέχει τη σταθερότητά και τα πρότυπα ανάλυσης των SNPs. Υπάρχει ο πίνακας βασικοί όροι / λέξεις-κλειδιά για κάθε γονίδιο, και ένας πίνακας της βάσης δεδομένων PubMed όπου αντλούνται οι ταυτότητες για κάθε γονίδιο. Ο πίνακας KnowledgeNet περιέχει τις υψηλές αλληλεπιδράσεις ανά ζεύγος γονιδίων, και υπάρχουν επίσης τα υποψήφια γονίδια που σχετίζονται με μία ασθένεια. Ορισμένοι άλλοι πίνακες που συνδέονται με τον πίνακα γονιδίων είναι: ο πίνακας Transcript (RefSeq mRNAs), ο πίνακας πρωτεϊνών (RefSeq πρωτεΐνες), ο πίνακας με το φαινότυπο και τις ασθένειες (NCBI-OMIM-HGMD), ο πίνακας πειραματικών δεδομένων (Bioscience mouse knockout), ο πίνακας μονοπατιών (KEGG), ο πίνακας πρωτεϊνικών αλληλεπιδράσεων (BIND) και ο πίνακας λειτουργίας των πρωτεϊνών (GO).

Εικόνα 1.1.5.6.1: Αναλυτικός σχήμα περιγραφής τρόπου συνδυασμού δεδομένων στον αλγόριθμο SNPs3D.



1.1.5.7 Γραφική διεπαφή KnowledgeNet του αλγόριθμου SNPs3D.

Η γραφική διεπαφή που δημιουργεί ο αλγόριθμος SNPs3d για την εμφάνιση των σχέσεων μεταξύ των γονιδίων βασίζεται σε κώδικα Java. Οι κόμβοι στο γράφημα αντικατοπτρίζουν τα γονίδια και οι συσχετισμοί μεταξύ των γονιδίων τα άκρα. Επιλέγοντας ένα σύνδεσμο οδηγούμαστε σε πιο λεπτομερείς πληροφορίες. Επιλέγοντας το σύμβολο που παρουσιάζει τα γονίδια μας μεταφέρει στα γονιδίου που εμπλέκονται στην ασθένεια. Τα υποσύνολα των γονιδίων που περιέχουν ένα ή περισσότερα SNPs σε συχνότητες πληθυσμού πάνω από κάποια όριο, μπορεί να επισημανθούν. Τα πρώτα 300 υποψήφια γονίδια που σχετίζονται με την ασθένεια εμφανίζονται στη γραφική διεπαφή. Αυτά είναι τα γονίδια που είναι πιο έντονα συνδεδεμένα με την ασθένεια. Το κατώτατο όριο για την εμφάνιση των συνδέσμων μεταξύ των γονιδίων είναι ρυθμιζόμενο ώστε να παρουσιάζονται τα πιο στενά συνδεδεμένα γονίδια.

Οι σύνδεσμοι μπορούν επίσης να βασίζονται σε KEGG συνδέσεις ή απευθείας πληροφορίες για τις αλληλεπιδράσεις μεταξύ των πρωτεϊνών, που προέρχονται από την βάση δεδομένων BIND[59]. Αριστερό κλικ σε ένα γονίδιο μας παραπέμπει άμεσα σε όλες τις συγκεκριμένες πληροφορίες του γονιδίου, συμπεριλαμβανομένων των SNP αναλύσεων, των μεθόδων προφίλ, και των NCBI περιλήψεων. Όλοι οι κατάλογοι με τα γονίδια μπορούν να υποστούν επεξεργασία. Ένα σημαντικό χαρακτηριστικό είναι η δυνατότητα να ξανασχεδιάσουμε το γράφημα, χρησιμοποιώντας ένα επιλεγμένο κόμβο, για το νέο κέντρο. Ένα μενού παρέχει μια λίστα με όλα τα γονίδια που εμφανίζονται, καθώς και κάθε γονίδιο μπορεί να επισημανθεί στο δίκτυο μέσω του εν λόγω καταλόγου.

1.1.5.8 Ανάλυση των SNPs σε κάθε ανθρώπινο γονίδιο.

Μια βασική λειτουργία του αλγόριθμου SNPs3D είναι ότι μας παρέχει ένα τρόπο για τον εντοπισμό των μη συνώνυμων SNPs που ενδέχεται να έχουν επιβλαβή επίδραση στη μοριακή λειτουργία in vivo, έτσι αυτά μπορούν να συμπεριληφθούν στις μελέτες σύνδεσης. Μία ανάλυση των πιθανών λειτουργικών επιπτώσεων όλων των μη συνώνυμων ανθρώπινων γενετικών παραλλαγών προβλέπονται από δεδομένα των HGMD και dbSNP βάσεων δεδομένων, χρησιμοποιώντας προηγούμενες μεθόδους[71].

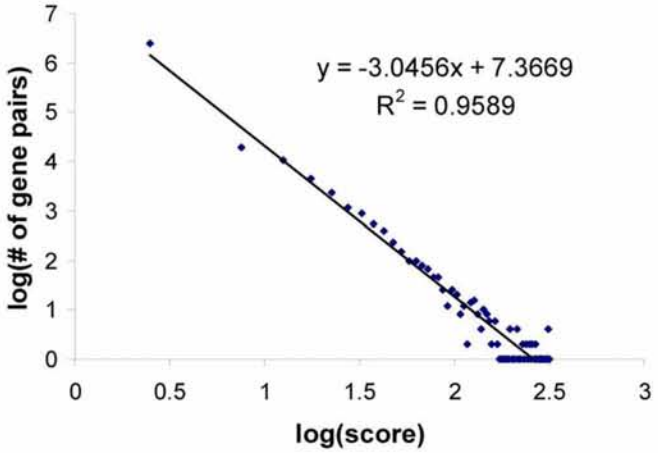
Η διαθεσιμότητα των πειραματικών δομών είναι επαρκώς ακριβής για την κάλυψη των ορίων σε μοντέλα δομής αφού το 37% των μονογονιδιακών παραλλαγών τις κάθε ασθένειας που περιέχονται στην HGMD βάση δεδομένων και ένα άλλο 10% στην dbSNP.

Στον αλγόριθμο SNPs3D χρησιμοποιείται μια τεχνική μηχανικής μάθησης διανυσματικών μηχανών στήριξης (SVM), ούτως ώστε να κρίνεται κάθε SNP ως επιβλαβές ή μη επιβλαβές. Οι διανυσματικές μηχανές στήριξης είναι εκπαιδευμένες στα δεδομένα μονογονιδιακών ασθενειών, έτσι ώστε ο ορισμός των επιβλαβών να είναι επιτυχώς επιβλαβή για την λειτουργία της πρωτεΐνης ώστε να είναι συνεπές με τα μονογονιδιακά αποτελέσματα της νόσου. Η συγκριτική αξιολόγηση απέδωσε εσφαλμένα θετικά 15% και ψευδώς αρνητικά 26% για 10% ευαισθησία της μεθόδου και 20% αντίστοιχα για τις μεθόδους προφίλ-ακολουθίας. Το υψηλότερο ψευδώς αρνητικό ποσοστό για τη σταθερότητα στην μέθοδο αυτή αντικατοπτρίζει το γεγονός ότι μόνο οι επιπτώσεις της σταθερότητας στην in vivo λειτουργία περιλαμβάνονται. Περίπου το 30% των μη συνώνυμων SNPs στην dbSNP βάση δεδομένων έχουν αποδοθεί ως επιβλαβές. Πολύ λίγες από τις περιπτώσεις των dbSNP δεδομένων είναι γνωστό ότι σχετίζονται με μονογονιδιακές ασθένειες και έτσι τα περισσότερα από αυτά τα επιβλαβή είναι υποψήφια για το ότι συμβάλουν στα πολύπλοκα χαρακτηριστικά της νόσου.

1.1.5.9 Σχέσεις μεταξύ γονιδίων.

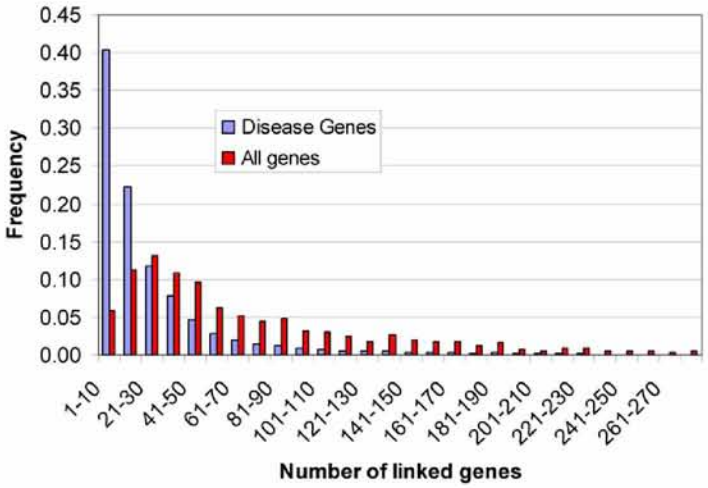
Η έννοια του επικαλυπτόμενου προφίλ χρησιμοποιήθηκε για να ιεραρχήσει τις σχέσεις μεταξύ όλων των ζευγών από ανθρώπινα γονίδια στην NCBI[37] βάση δεδομένων.

Εικόνα 1.1.5.9.1: Κατανομή των βαθμών ανά ζεύγη.



Όταν μόνο τα ζεύγη των γονιδίων εμπλέκονται στις φυσικές αλληλεπιδράσεις που περιλαμβάνονται στη βάση δεδομένων BIND[59], η διάμεσος και η μέση τιμή είναι δραματικά υψηλότερη, στα 3,2 και 9,0 αντίστοιχα.

Εικόνα 1.1.5.9.2: Στην γραφική παράσταση παρουσιάζεται η κατανομή του αριθμού των γονιδίου και πώς συνδέονται με τις μονογονιδιακές ασθένειες.



Η Εικόνα 1.1.5.9.2 δείχνει την κατανομή του αριθμού των γονιδίων και πώς συνδέονται με τις μονογονιδιακές ασθένειες (που ορίζεται με την συμπερίληψη της HGMD βάσης δεδομένων) , για όλα τα γονίδια. Τα γονίδια που σχετίζονται με ασθένειες έχουν την τάση να συνδέονται με περισσότερα γονίδια από τα μη συσχετισμένα με ασθένειες γονίδια, γεγονός που αντικατοπτρίζει ότι είναι συνήθως καλά μελετημένα και έχουν τοποθετηθεί σε ένα πλαίσιο στο δίκτυο.

1.1.5.10 Ανάλυση των γονιδίων στην διεπαφή KnowledgeNet για να διερευνήσουμε τις SNP φαινοτυπικές σχέσεις.

Τα SNPs στην Εικόνα 1.1.5.10.1 ταξινομούνται ως πολύ επιβλαβή με βάση τις πρωτεΐνες, και τα γονίδια που σχετίζονται με ασθένειες. Ωστόσο, κανένα από αυτά τα είναι SNPs δεν σχετίζεται με βάση το φαινότυπου του στις ασθένειες. Η διεπαφή KnowledgeNet μπορεί να χρησιμοποιηθεί για τη διερεύνηση της πολύπλοκης σχέσης μεταξύ της επίδρασης αυτών των SNPs σε επίπεδο λειτουργίας των πρωτεϊνών και φαινότυπου της νόσου, μέσω του δικτύου. Θεωρούμε ένα ζεύγος γονιδίων με επιβλαβή SNPs, σελεκτίνης E και σελεκτίνη P. Η πλαϊνή στήλη στην σελίδα ανάλυσης των SNP παρέχει άμεση πρόσβαση σε ένα ευρύ φάσμα πληροφοριών σχετικά με το θέμα αυτό,

συμπεριλαμβανομένης OMIM βάσης δεδομένων, GO βάσης δεδομένων, και πειραματικών δεδομένων αλλά και παρέχει ειδική ιστολογική έκφραση των δεδομένων, και των σχετικών περιλήψεων. Το διάγραμμα γονιδίου δημιουργεί ένα παράθυρο Java που εμφανίζει τις σχέσεις μεταξύ γονίδιου- γονίδιου με επίκεντρο τα πιο υψηλά υποψήφια γονίδια που σχετίζεται με την ασθένεια.

Εικόνα 1.1.5.10.1 Παράδειγμα με SNPs που είναι υποψήφια να σχετίζονται με μία ασθένεια.

	refseq accession	snp	snp_id	svm profile	svm structure	molecular effect	model	frequency
SELE	NP_000441	C130W	5360	-1.89	-1.06	OverPacking Breakage of a disulfide bond;		0.02
SELP	NP_002996	G179R	3917718	-0.81	-1.46	OverPacking Backbone Strain;		0.02
SELL	NP_000646	P213S	4987310	-0.36				0.21
VCAM1	NP_001069	S318F	3783611	-1.31				0.03
VCAM1	NP_001069	G413A	3783613	-0.96				0.08
VCAM1	NP_542413	I624L	3783615	-0.68				0.06

Υπάρχουν τρία μοναδικά χαρακτηριστικά γνωρίσματα του αλγόριθμου SNPs3D. Πρώτον, είναι σχεδιασμένος ειδικά για την ανάλυση της σχέσης μεταξύ SNPs και των ασθενειών (Εικόνα 1.1.5.10.2). Δεύτερον, κατασκευάζει δίκτυα από γονίδια με βάση τα παράγωγα των εννοιολογικών συσχετίσεων από τις δημοσιεύσεις και όχι από πειραματικά δεδομένα. Τρίτον, ενσωματώνει συνδέσμους σε όλες τις διαθέσιμες πηγές με σχετικές πληροφορίες, δίνοντας στο χρήστη την εύκολη πρόσβαση για τα βασικά δεδομένα και τις δημοσιεύσεις που αφορούν το κάθε γονίδιο.

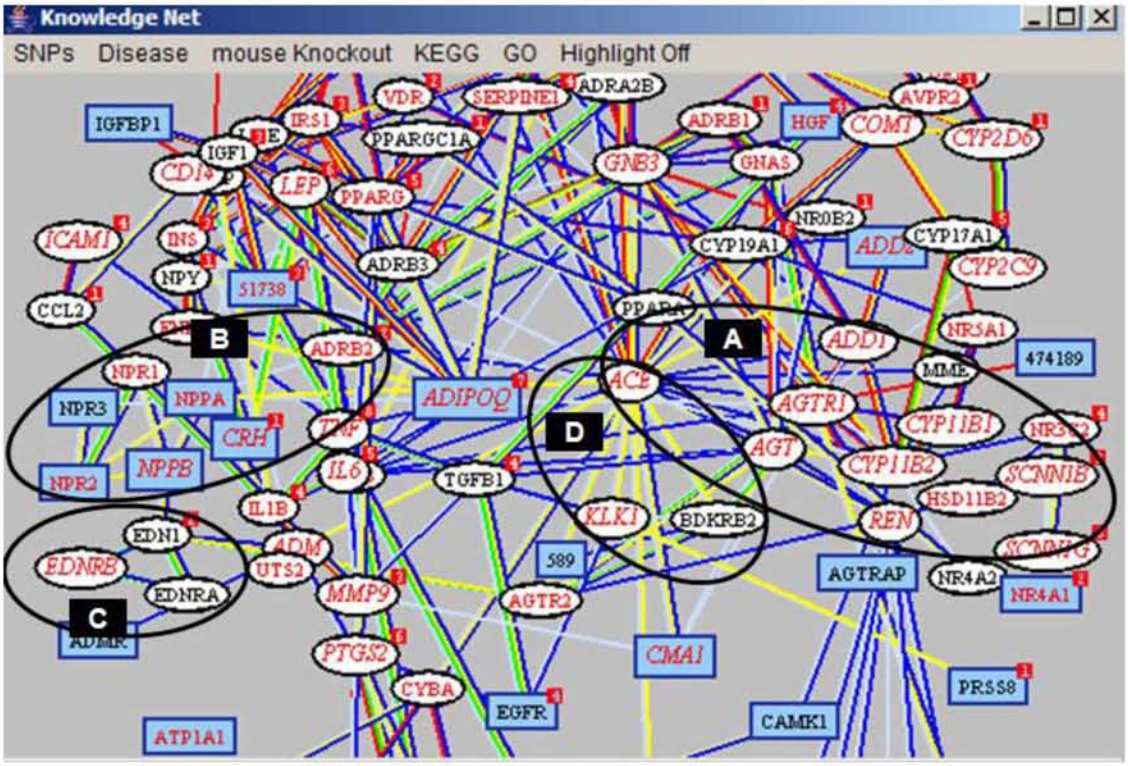
Ένα δίκτυο συνδέσεων κατασκευάζεται μεταξύ των γονιδίων και βασίζεται στο πόσο έντονα είναι συνδεδεμένα τα γονίδια με τις δημοσιεύεις. Υπάρχουν δύο πλεονεκτήματα σε αυτή την προσέγγιση. Η σχετική σύνδεση μεταξύ πρωτεϊνών μπορεί να μην είναι φυσική. Για παράδειγμα, τα γονίδια που εμπλέκονται στην ίδια ασθένεια δεν μπορούν να αλληλεπιδρούν άμεσα. Επίσης δεν μπορούν να είναι στο ίδιο τοπικό μονοπάτι, αλλά και ούτε να αλληλεπιδρούν από την άποψη ότι επηρεάζουν την ευαισθησία της ασθένειας. Έτσι η KnowledgeNet διευρύνει τις υπάρχουσες οδούς περιγραφών, συνδέοντας τα γονίδια με εννοιολογικές συσχετίσεις. Με βάση τα

αποτελέσματα δείχνουν πως όλο αυτό λειτουργεί στην πράξη. Η ανάλυση των μη συνώνυμων SNPs οδηγεί στη διαπίστωση ότι φαίνεται να είναι επιβλαβή για την λειτουργία των πρωτεϊνών, αλλά δεν οδηγούν κατευθείαν σε μια ασθένεια με το φαινότυπο.

Η επιθεώρηση των ομόλογων SNPs στην γραφική διεπαφή KnowledgeNet προτείνει ένα λειτουργικό δίκτυο που παρέχει επίπεδο ευρωστίας και ελέγχει με πειραματικά δεδομένα τα στοιχεία από τα πρότυπα έκφρασης που χρησιμοποιούνται ως παράμετροι στην όλη διαδικασία. Όπως συμβαίνει συχνά, μερικά από τα γονίδια που σχετίζονται με την νόσο, η ευαισθησία τους είναι περίπλοκη, και οι διαθέσιμες πληροφορίες είναι μερικές φορές αντιφατικές.

Τα πλεονεκτήματα της προσέγγισης είναι ότι δημιουργούνται προφίλ για κάθε γονίδιο με περισσότερες λεπτομέρειες αλλά και με τις σχετικές λέξεις και όρους από τις δημοσιεύσεις της PubMed βάσης δεδομένων. Με τη σειρά τους, η επικάλυψη των προφίλ χρησιμοποιείται για την ταυτοποίηση των σχέσεων γονιδίων με γονίδιο.

Εικόνα 1.1.5.10.2: Γραφική αναπαράσταση υποψήφιων γονιδίων από την διεπαφή KnowledgeNet.



Στην πράξη, η διαδικασία προβλέπει λογικά αποτελέσματα, αλλά δεν υπάρχει κανένας τρόπος να είναι ικανός να κάνει συγκριτική αξιολόγηση αυτών των αποτελεσμάτων που δημιουργούνται στα δίκτυα. Η μέθοδος σφάλλει κατά καιρούς από την πλευρά της υπερβολικής ένταξης καινούργιων και μη αξιόπιστων συγγραμμάτων. Ομοίως, σχέσεις μεταξύ γονιδίων μπορεί μερικές φορές να βασίζονται σε μη συναφείς παράγοντες.

1.1.6 Αλγόριθμος FitSNPs.

Ο αλγόριθμος FitSNPs[85], χρησιμοποιώντας τον τεράστιο πλούτο από πρότυπα έκφρασης γονιδίων που παρέχονται ελεύθερα, ιεραρχεί και ταξινομεί τα υποψήφια γονίδια που σχετίζονται με ασθένειες και αιτιατές παραλλαγές. Σε αυτό τον αλγόριθμο αναλύθηκαν όλες οι διαθέσιμες ανθρώπινες μελέτες μικροσυστοιχιών από την βάση δεδομένων GEO[86] και υπολογίστηκε η συχνότητα της διαφορετικής έκφρασης του κάθε γονίδιου. Όσο πιο συχνά ένα γονίδιο ήταν διαφορετικά εκφρασμένο, τόσο πιθανότερο ήταν ότι συνδέεται με ασθένεια που σχετίζεται με γενετικές παραλλαγές. Με βάση αυτή την ανακάλυψη, δημιουργήθηκε ένας κατάλογος με όλα τα λειτουργικά παρεμβαλλόμενα SNPs (fitSNPs) από την διαφορική έκφραση των γονιδίων. Χρησιμοποιώντας τα λειτουργικά παρεμβαλλόμενα SNPs αποδείχθηκε ότι ταιριάζουν και ότι θα μπορούσαν να είχαν χρησιμοποιηθεί για να δοθεί ιεράρχηση με επιτυχία στα υποψήφια γονίδια.

1.1.6.1 Τρόπος βαθμολόγησης υποψήφιων γονιδίων με βάση τον αλγόριθμο FITSNPs.

Όλα τα ανθρώπινα γονίδια ταξινομούνται με βάση το κριτήριο «τα πιο υψηλά πιθανά ευπαθή γονίδια σχετίζονται με κάποιες γενετικές παραλλαγές σε συνδυασμό της πειραματικής επιβεβαίωσης της νόσου». Η πιθανότητα αυτή υπολογίζεται από συντελεστές με βάση την διαφορετική έκφραση τους (DER). Όσο υψηλότερη είναι η τιμή DER σε ένα γονίδιο, τόσο πιο πιθανό είναι ότι θα σχετίζεται με μία ασθένεια που συνδέεται με κάποιες γενετικές παραλλαγές. Επιλέγοντας την τιμή DER, μπορούμε να ανακτήσουμε όλες τις σχετικές μελέτες από μικροσυστοιχίες, όπου τα υψηλά υποψήφια γονίδια (top genes) εκφράζονται διαφορετικά. Επιλέγοντας την τιμή GAD, μπορούμε να ανακτήσουμε τις πειραματικές αποδείξεις με τις οποίες συνδέεται η νόσος.

Ο κάθε χρήστης μπορεί να ιεραρχήσει τα υποψήφια ευπαθή γονίδια εισάγοντας απλά την περιοχή τους (locus) αφού ο αλγόριθμος αυτός εκτελεί αναζητήσεις για την άγνωστη μοριακή τους σύσταση από την βάση δεδομένων OMIM[24]. Μέσα από πειράματα που εκτελέστηκαν στον αλγόριθμο αυτό, κατέληξαν στο συμπέρασμα ότι κάποια υψηλά διαφορετικά εκφρασμένα γονίδια δεν έχουν ακόμη συσχετιστεί με κάποια ασθένεια. Με αυτά τα αποτελέσματα δίνετε η δυνατότητα στους ερευνητές να ασχοληθούν με ενδιαφέρουσες έρευνες.

1.1.6.2 Αποτελέσματα αλγορίθμου FITSNPs.

Τα υψηλά διαφορετικά εκφρασμένα γονίδια είναι πιο πιθανό να σχετίζονται με ασθένειες με βάση τις διάφορες γενετικές παραλλαγές που παρουσιάζονται στην κάθε ασθένεια. Προκειμένου να καθοριστεί αν τα διαφορετικά εκφρασμένα γονίδια σχετίζονται γενετικός με τη νόσο, κατεβάστηκαν όλα τα ανθρώπινα σύνολα δεδομένων (476) από την βάση δεδομένων GEO για να χρησιμοποιηθούν στα ανθρώπινα εκφρασμένα γονίδια που έχουν τεθεί. Αυτά τα σύνολα δεδομένων της GEO βάσης δεδομένων, αφορούν τις ομάδες μικροσυστοιχιών που δημιουργούνται από πειραματικά δεδομένα. Χρησιμοποιώντας τον αλγόριθμο AILUN[87], ακολούθως σχολιάστηκαν τα σύνολα με πληροφορίες από την NCBI βάση δεδομένων. Γενικά στον αλγόριθμο FITSNPs διενεργήθηκαν 4.877 συγκρίσεις συνόλων χρησιμοποιώντας σημασιολογική ανάλυση μικροσυστοιχιών με την μέθοδο SAM[88] και έτσι παράχθηκε μια λίστα με 19.879 γονίδια που ήταν διαφορικά εκφράζονται με ποσοστό (q) ψευδός θετικών προβλέψεων μικρότερο του 0,05 σε ένα ή περισσότερα πειράματα.

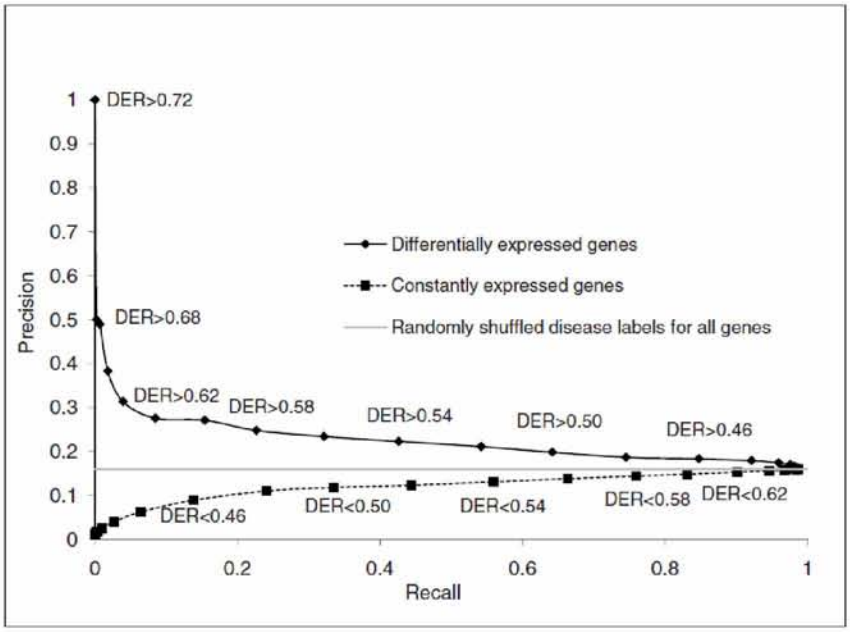
Ακολούθως δημιούργησαν μια λίστα με τα επιμελημένα ανθρώπινα γονίδια ανακτώντας πληροφορίες από τις βάσεις δεδομένων GAD[56] και HGMD[51], με αποτέλεσμα να έχουμε μια λίστα με 3.221 γονίδια που σχετίζονται με γενετικές παραλλαγές ασθενειών. Συγκρίνοντας την λίστα των διαφορικά εκφρασμένων γονιδίων με την λίστα των ευπαθή γονιδίων που σχετίζονται με γενετικές παραλλαγές, διαπιστώθηκε ότι το 99% των γονιδίων που σχετίζονται με ασθένειες ήταν διαφορικά εκφρασμένα σε ένα ή περισσότερα σύνολα δεδομένων της βάσης δεδομένων GEO, με ειδικότητα 14% . Η πιθανότητα ύπαρξης γενετικών παραλλαγών που να συνδέονται με τη νόσο ήταν 12 φορές υψηλότερη μεταξύ των διαφορικά εκφρασμένων γονιδίων σε σχέση με τα συνεχή εκφρασμένα γονίδια, ενώ η πιθανότητα της ύπαρξης μίας μη συνώνυμης SNP κωδικοποίησης ήταν 1,6 φορές υψηλότερη μεταξύ των διαφορικά εκφρασμένων γονιδίων σε σχέση με τα συνεχή εκφραζόμενα γονίδια.

1.1.6.3 Αποτελέσματα αλγορίθμου FITSNPs / Έκφραση DER.

Για καλύτερο χαρακτηρισμό της σχέσης μεταξύ της DNA διακύμανσης και έκφρασης σε όλα τα ανθρώπινα γονίδια, ελέγχθηκε αν τα διαφορικά εκφρασμένα γονίδια σε πολλαπλές μελέτες μικροσυστοιχιών έχουν περισσότερες πιθανότητες να συνδέονται με ασθένειες που σχετίζεται με γενετικές παραλλαγές. Για κάθε γονίδιο, η αναλογία

διαφορικής έκφρασης (DER) υπολογίστηκε ως ο αριθμός του συνόλου δεδομένων από την GEO βάση δεδομένων στην οποία είχε διαφορικά εκφρασμένα γονίδια (q value ≤ 0,05) προς του ολικού συνολικού αριθμού δεδομένων της βάσης δεδομένων GEO. Ο υπολογισμός που έγινε περιορίζεται στα γονίδια που μετρήθηκαν αφού τουλάχιστον 5% του συνόλου είναι σύνολα δεδομένων της GEO βάσης δεδομένων.

Εικόνα 1.1.6.3.1: Διακύμανση της τιμής DER.

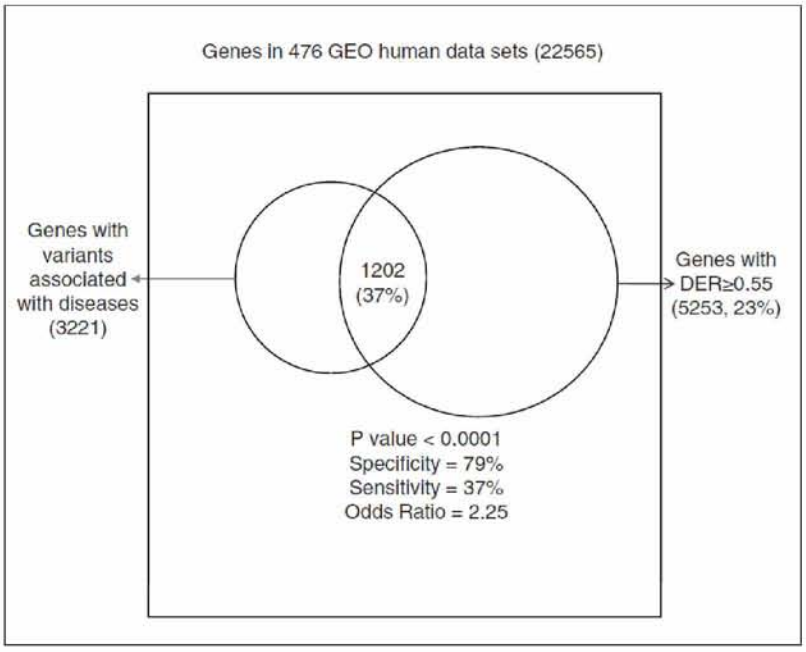


Η ακρίβεια των εκ νέου ευρισκόμενων γονιδίων που σχετίζονταν με τη νόσο ήταν 16% για τα γονίδια με DER μεγαλύτερο από 0. Το ποσοστό αυτό βελτιώθηκε σταδιακά στο 28%, όταν η τιμή DER ήταν μεγαλύτερη από 0,62, και στη συνέχεια αυξήθηκε δραματικά σε ποσοστό 100%, όταν η τιμή DER ήταν μεγαλύτερη από 0,72 (Εικόνα 1.1.6.3.1). Τα περισσότερα σύνολα δεδομένων της GEO βάσης δεδομένων στα οποία ένα γονίδιο ήταν διαρκώς εκφρασμένο, ήταν μεγαλύτερο πιθανό να συνδέεται με ασθένειες που σχετίζονται με γενετικές παραλλαγές.

Σε ένα δέκτη καμπύλης κατασκευασμένος έτσι ώστε να ανακαλύψουμε ξανά τα γονίδια που σχετίζονται με την ασθένεια χρησιμοποιώντας τις τιμές DER, η τιμή DER ≥0,55 παρουσίασε τις καλύτερες επιδόσεις, με ειδικότητα 79% και 37% ευαισθησία. Όπως φαίνεται στην Εικόνα 1.1.6.3.2, τα γονίδια με DER ≥0,55 είχαν 2,25 φορές μεγαλύτερη πιθανότητα να σχετίζονται με ασθένειες που σχετίζεται με γενετικές παραλλαγές σε σχέση με άλλες (P <0,0001, έλεγχος Fisher). Διαφοροποιώντας το

κατώτατο όριο, είχαν 56% ειδικότητα και 65% ευαισθησία με τιμή DER ≥0,50, και 93% ειδικότητα και 16% ευαισθησία με τιμή DER ≥0,60.(Εικόνα 1.1.3.3.2)

Εικόνα 1.1.6.3.2: Απόδοση των επαληθευμένων γονιδίων που σχετίζονται με τη νόσο με βάση την τιμή DER.



1.1.6.4 Συγκρίσει των τιμών DER μεταξύ των διαφόρων τύπων γονιδίων που σχετίζονται με ασθένειες.

Η επιτυχία των μελετών επικύρωσης καταδεικνύουν ότι ο αλγόριθμος fitSNPs θα μπορούσε να χρησιμοποιηθεί όχι μόνο να δώσει προτεραιότητα σε διαφορετικές περιοχές ενδιαφέροντος μέσα σε όλο το γονιδίωμα αλλά και για να επιλέγει γονίδια από κάθε γενετική περιοχή. Πριν την εφαρμογή του fitSNPs σε όλες τις νόσους, ένα σημαντικό ζήτημα είναι αν τα γονίδια που σχετίζονται με τους διαφορετικούς τύπους ασθενειών έχουν διαφορετικές τιμές DER. Έτσι κατεβάστηκαν όλες οι λίστες τα γονίδια που σχετίζονται με Μεντελικές ασθένειες (Ασθένειες που προκαλούνται από μία μόνο μετάλλαξη), λίστες ασθενειών για γονίδια πολυπαραγοντικής αιτιολογίας, και λίστες με γονίδια καρκίνου[89]

Όπως φαίνεται στον Πίνακα 1.1.6.4.1, δεν παρατηρήθηκε σημαντική διαφορά της τιμής DER στα γονίδια μεταξύ των Μεντελικών ασθενειών και ασθενειών πολυπαραγοντικής αιτιολογίας (0,53 έναντι 0,54 P = 0,2, t-test). Όμως τα καρκινικά γονίδια εκτίθενται σε σημαντικά υψηλότερες τιμές DER (0,56) σε σχέση με όλα τα

Μεντελικά γονίδια(P <0,0001, t-test) και τα γονίδια που σχετίζονται με ασθένειες πολυπαραγοντικής αιτιολογίας (P = 0,001, t-test). Επιπλέον, όλοι οι τύποι των γονιδίων που σχετίζονται με ασθένειες εκτίθενται σε σημαντικά υψηλότερες τιμές από ότι οι τιμές DER των γονιδίων που δεν σχετίζονται με ασθένειες (P <0,0001, t-test). Τα ευρήματα αυτά υποδηλώνουν ότι ο αλγόριθμος fitSNPs θα μπορούσε να χρησιμοποιηθεί για να δώσει προτεραιότητα σε γονίδια που σχετίζονται σε ασθένειες και για τους δύο τύπους Μεντελικών ασθενειών αλλά και στις ασθένειες πολυπαραγοντικής αιτιολογίας. Επίσης είναι αρκετά αποτελεσματικός στην ιεράρχηση γονιδίων που σχετίζονται με το καρκίνο.

Πίνακας 1.1.6.4.1: Τιμές DER συγκρίνοντας μεταξύ Μεντελικών, πολυπαραγοντικής αιτιολογίας, καρκινικών, διαφόρων τύπων, γονιδίων που σχετίζονται με ασθένειες και των γονιδίων που δεν σχετίζονται με ασθένειες.

P value	Μεντελικά (μέσος = 0.53, n=931)	Πολυπαραγοντικής αιτιολογίας (μέσος = 0.54, n=70)	Καρκινικά (μέσος = 0.56, n=324)	Διαφόρων τύπων (μέσος = 0.53, n=3.178)	Μη ευπαθή (μέσος = 0.50, n=16.698)
Μεντελικά γονίδια		0.2	<0.0001	0.4	<0.0001
Πολυπαραγοντικής αιτιολογίας γονίδια			0.001	0.3	<0.0001
Καρκινικά γονίδια				<0.0001	<0.0001
Διαφόρων τύπων ασθενειών γονίδια					<0.0001
Μη ευπαθή γονίδια					

Ο αλγόριθμος FitSNPs προβλέπει τα υποψήφια γονίδια που σχετίζονται με ασθένειες με βάση τις πληροφορίες που παίρνει από τη βάση δεδομένων OMIM για τις περιοχές ενδιαφέροντος που αναζητεί με άγνωστη μοριακή βάση. Ο FitSNPs μπορεί να χρησιμοποιηθεί για να δώσει προτεραιότητα σε γονίδια που σχετίζονται με πολλά είδη ασθενειών από όλο το φάσμα του ευρέως γονιδιώματος, αλλά και για να προβλέψει γενετικές παραλλαγές ασθενειών για τα γονίδια με υψηλές τιμές DER. Υπάρχουν 5.253 ανθρώπινα γονίδια με τιμή DER ≥ 0,55. Από αυτά, το 23% ανήκει σε γνωστές γενετικές παραλλαγές ασθενειών σύμφωνα με την βάση δεδομένων GAD και HGMD. Τα υπόλοιπα 4.052 γονίδια δεν έχουν ακόμη αποδειχθεί να συνδέονται με οποιαδήποτε ασθένεια μετάλλαξης ή πολυπαραγοντικής αιτιολογίας, καθιστώντας τα ελπιδοφόρα. Για την συστηματική πρόβλεψη συσχετισμένων νόσων, αναζητήθηκαν από την βάση δεδομένων OMIM γονίδια και διαπιστώθηκε ότι οι 830 ασθένειες και σύνδρομα έχουν συνδεθεί με τις κυτταρογενετικές περιοχές, αλλά όχι με συγκεκριμένα γονίδια. Από αυτές

τις κυτταρογενετικές περιοχές , έχουν προβλεφτεί 3.331 υψηλά διαφορετικά εκφρασμένα γονίδια με τιμή DER ≥ 0,55 σε 610 ασθένειες. Από αυτή την ομάδα, τα 2.586 γονίδια δεν συνδέονται με κάποια ασθένειας, σύμφωνα με της βάσης δεδομένων GAD και HGMD, που είχαν προβλέψει να συνδέονται με 597 ασθένειες.

1.1.7 Αλγόριθμος PosMed.

Ο αλγόριθμος PosMed [90] εκτελεί ιεράρχηση υποψήφιων γονιδίων που σχετίζονται με ασθένειες με την μέθοδο «positional cloning». Η μέθοδος αυτή εκτελεί ταυτοποίηση του γονιδίου με τα χαρακτηριστικά μία ασθένειας με βάση τους σχολιασμούς που το χαρακτηρίζουν από τις διάφορες δημοσιεύσεις. Η διαδικασία εξόρυξης δεδομένων γίνεται από την βάση αναζήτησης του αλγόριθμου PosMed, GRASE ένα διαδικτυακό ταξινομητή ο οποίος χρησιμοποιεί ένα συγκεκριμένο τρόπο καθορισμού των δεδομένων αλλά και τις μεταξύ τους συνδέσεις, πολύ κοντινό με την τεχνική των νευρωνικών δικτύων. Ο ταξινομητής αυτός περιλαμβάνει τεκμηριωμένους νευρώνες «documentrons» που αντιπροσωπεύουν κάθε έγγραφο που προέρχεται από τις βάσεις δεδομένων MEDLINE [91] και OMIM [24].

Αρχικά δίνονται στον ταξινομητή οι λέξεις κλειδιά που σχετίζονται με την ασθένεια που αναζητούμε. Εκτελείται μία πλήρης γενική αναζήτηση για τον κάθε τεκμηριωμένο νευρώνα «documentron» και ακολούθως μαζεύονται τα δεδομένα για κάθε ένα από αυτούς τους τεχνητούς νευρώνες πρώτου επιπέδου. Στη συνέχεια υπολογίζεται η στατιστική σημασία των μεταξύ τους συνδέσεων που βρέθηκε για κάθε τεκμηριωμένο νευρώνα «documentrons» και δημιουργούνται οι τεχνητοί νευρώνες δεύτερου επιπέδου που αντιπροσωπεύουν κάθε γονίδιο.

Όταν ένα χρωμοσωμικό μεσοδιάστημα (s) ορίζεται, ο αλγόριθμος PosMed διερευνά το δεύτερο και το τρίτο επίπεδο τεχνητών νευρώνων που εκπροσωπούν τα γονίδια με το χρωμοσωμικό διάστημα εξελίσσοντας την συνδυασμένη σημαντικότητα των συνδέσεων που έχει ο κάθε σημαδεμένος τεκμηριωμένος νευρώνας με τα γονίδια (σημαδεμένος ορίζεται αυτός για τον οποίο έχει βρεθεί ένας συγκεκριμένος αριθμός δημοσιεύσεων).

Ο αλγόριθμος PosMed, επομένως είναι ένα ισχυρό εργαλείο που κατέχει άμεσα τα υποψήφια γονίδια με την χρήση της φαινοτυπικής σύνδεσης των λέξεις-κλειδιών με τα γονίδια μέσω των συνδέσεων που αντιπροσωπεύουν τις αλληλεπιδράσεις γονιδίων με τα γονίδια, αλλά και άλλων βιολογικών αλληλεπιδράσεων και ορθολογικών δεδομένων. Ο αλγόριθμος PosMed χρησιμοποιώντας τις ορθολογικές συνδέσεις, κάνει πολύ πιο εύκολα την κατάταξη των ανθρώπινων γονιδίων με βάση τα στοιχεία που βρέθηκαν σε άλλα είδη μοντέλων όπως το ποντίκι.

Η ανάλυση σύνδεσης γενικά χρησιμοποιείται για τον προσδιορισμό και ταυτοποίηση των γονιδίων πού έχουν ένα συγκεκριμένο φαινότυπο ή μία γενετική ανωμαλία, και προτείνει χρωμοσωμικές περιοχές που περιέχουν αρκετά μεγάλο αριθμό γονιδίων που είναι υποψήφια να σχετίζονται για την εκδήλωση μίας ασθένειας . Αρχικά πριν την περεταίρω ανάλυση του αλγόριθμου PosMed είναι αναγκαίο να δοθεί προτεραιότητα στα υποψήφια γονίδια που πιθανός να σχετίζονται με την ασθένεια που αναζητούμε με τη χρήση όλων των βιολογικών γνώσεων και γενικά δεδομένων που μπορούμε να έχουμε.

Μεγάλη πρόκληση του αλγόριθμου PosMed είναι να αναπτύξει ένα διαδικτυακό εργαλείο που να μπορεί να προτείνει άμεσα και αξιόπιστα διάφορα γονίδια που πιθανός να σχετίζονται με μία ασθένεια. Ο αλγόριθμος PosMed έχει υλοποιήσει αυτό το διαδικτυακό εργαλείο και του έδωσε την ονομασία GRASE, όπου είναι ένας διαδικτυακός ταξινομητής γενικής και ειδικής (ποντίκι) συσχετίσεις μελετών. Ο ταξινομητής GRASE υλοποιείται σε GRASQL[92] όπου είναι μία καλή γλώσσα προγραμματισμού για στατιστική ανάλυση δεδομένων που ανακτώνται από την γλώσσα SPARGL[93] όπου αυτή αναζητά πληροφορίες για τα εξεταζόμενα γονίδια μας από τις βάσεις δεδομένων που βρίσκονται στο διαδίκτυο.

Πίνακας 1.1.6.1: Σύνολα δεδομένων που συμπεριλαμβάνονται στην PosMed.

Έγραφα	Όνομα στην PosMed	Αριθμός εγγράφων	Αναφορά
MEDLINE	MEDLINE	17132801	[91, 94]
BRMM	Μεταλλαγμένο Ποντίκι	12911	[95]
OMIM	OMIM	19891	[24]
HsPPI	HsPPI	35731	[96]
AtPID	AtPID	44082	[97]
ATTED-II	Co-expressions	24418	[98]
REACTOME	REACTOME	10761	[99]
MouseGeneRecord	Mouse Gene Record	58768	[100]
RatGeneRecord	Rat Gene Record	36634	[101]
HumanGeneRecord	Human Gene Record	31459	[102]
ArabidopsisGeneRecord	Arabidopsis Gene Record	32041	[103]
MetaboliteRecord	Metabolite Record	18045	[104]
DrugRecord	Drug Record	1015	[95]
DiseaseRecord	Disease Record	1911	[95]
RIKENRescearcherRecord	Researcher record	6852	[95]
Σύνολο		17467320	

Μετά από πολλές μελέτες και βελτιώσεις του ταξινομητή GRASE ο ταξινομητής αυτός μπορεί να υπολογίσει αποτελεσματικά την στατιστική ιεράρχηση των υποψήφιων

γονιδίων που σχετίζονται με ασθένειες, βασιζόμενος στην βιβλιογραφία άνω των 17 εκατομμυρίων ιατρικών και βιολογικών εγγράφων σε πάρα πολύ γρήγορους ρυθμούς.

Τα διάφορα εργαλεία λογισμικού που έχουν αναπτυχθεί για τον καθορισμό της ιεραρχικής θέσης που θα παίρνει το κάθε υποψήφιο γονίδιο αλλά και για τον τρόπο ιεράρχησης τους βασίζονται στις λειτουργικές περιγραφές του κάθε γονιδίου, στα πρότυπα έκφρασης του, στις αλληλεπιδράσεις μεταξύ των πρωτεϊνών του και στα διάφορα χαρακτηριστικά που μπορούμε να εξάγουμε με βάση την ακολουθία του.[9, 105-109]

Τα έγγραφα και οι σχετικές πληροφορίες που αναζητούνται από τον αλγόριθμο PosMed περιέχουν δημοσιευμένες έρευνες που σχετίζονται με το γονίδιο ή την ασθένεια που αναζητούμε με βάση, σχολιασμούς γονιδιώματος, πληροφορίες φαινότυπου, αλληλεπιδράσεις μεταξύ πρωτεϊνών , co-expressions, ορθόλογα γονίδια και πληροφορίες μεταβολισμού (Πίνακας 1.1.6.1).

1.1.6.1 Αναπαράσταση των στατιστικών αναλύσεων των συσχετισμένων εγγράφων με βάση τα νευρωνικά δίκτυα.

Οι αναζητήσεις του αλγόριθμου PosMed εκτελούνται από τον ταξινομητή GRASE, μία μηχανή αναζήτησης που ανακτά στοιχεία δεδομένων σε ένα εξαιρετικά συνδεδεμένο δίκτυο από σημασιολογικές σχέσεις με βάση την αξιολόγηση των στατιστικών αναλύσεων.

- Αρχικά, εντοπίζονται τα γονίδια που σχετίζονται τις λέξεις-κλειδία που εισάγει ο χρήστης. Ο ταξινομητής GRASE εκτελεί αναζήτηση πλήρους κειμένου χρησιμοποιώντας τις λέξεις –κλειδιά και τα παρουσιάζει σε μορφή γράφου. Για τα κείμενα που συνδέονται με αυτές τις λέξεις κλειδιά, δημιουργεί τις κατάλληλες συνδέσεις σε μορφή μονοπατιών στο γράφο ανάμεσα στις δημοσιεύσεις/έγγραφα που βρίσκει και στα γονίδια που εξετάζει. Δηλαδή προσδιορίζει τα έγγραφα/δημοσιευσεις που έχουν την λέξη-κλειδί και δημιουργεί σε μορφή μονοπατιού στο γράφημα τη σημασιολογική σύνδεση της λέξης κλειδί με το έγγραφο.

- Για κάθε γονίδιο, δημιουργείται ένας πίνακας συνάφειας 2x2 με βάση το γράφημα που έχει δημιουργηθεί. Γεμίζει την πρώτη στήλη του πίνακα συνάφειας με τις λέξεις κλειδιά που ταιριάζουν με τα έγραφα και την δεύτερη με τις σχέσεις γονιδίου – εγγράφου/δημοσιεύσης ανάλογα με τα μονοπάτια που υπάρχουν στο γράφο. Για κάθε

ενδεχόμενο πίνακα συνάφειας υπολογίζεται το p-value χρησιμοποιώντας ένα στατιστικό έλεγχο όπως ο έλεγχος Fisher[110]. Η τιμή του p-value είναι μικρότερη όταν η συνάφεια μεταξύ των λέξεων-κλειδιών και των γονιδίων είναι πολύ μεγάλη. Η τιμή αυτή αποθηκεύεται και χρησιμοποιείται για την αξιολόγηση της συνάφειας μεταξύ των γονιδίων και των λέξεων-κλειδιών.

Για να εντοπίσει τα γονίδια που σχετίζονται με τα περαιτέρω γονίδια που αρχικά βρέθηκαν, ο ταξινομητής GRASE προβαίνει σε έρευνα για να ελέγξει κατά πόσο σχετίζονται μεταξύ τους τα γονίδια παίρνοντας ζεύγη με δυο γονίδια (γονίδιο1 – γονίδιο2) κατά την ίδια διαδικασία που εκτέλεσε πιο πάνω. Δηλαδή παράγει πίνακες συνάφειας 2x2 για κάθε γονίδιο βλέποντας την σχετική σύνδεση στο γράφημα που ταιριάζει πάνω η σημασιολογική σύνδεση εγγράφου/δημοσιεύσης με το γονίδιο και υπολογίζει το p-value. Όπως και πιο πάνω όσο μικρότερο είναι το p-value τόσο μεγαλύτερη είναι η συνάφεια μεταξύ των δύο γονιδίων με βάση των αριθμό των εγγράφων που μοιράζονται. Η τιμή αυτή χρησιμοποιείται για την αξιολόγηση της σχέσης μεταξύ των δύο γονιδίων.

Το συνολικό p-value υπολογίζεται με το συνδυασμό των προηγούμενων δύο p-value, το οποίο χρησιμοποιείται για να δείξει την στατιστική σημαντικότητα μεταξύ των λέξεων-κλειδιών και του γονιδίου2 μέσω γονιδίου1. Το συνολικό p-value υπολογίζεται ως εξής:

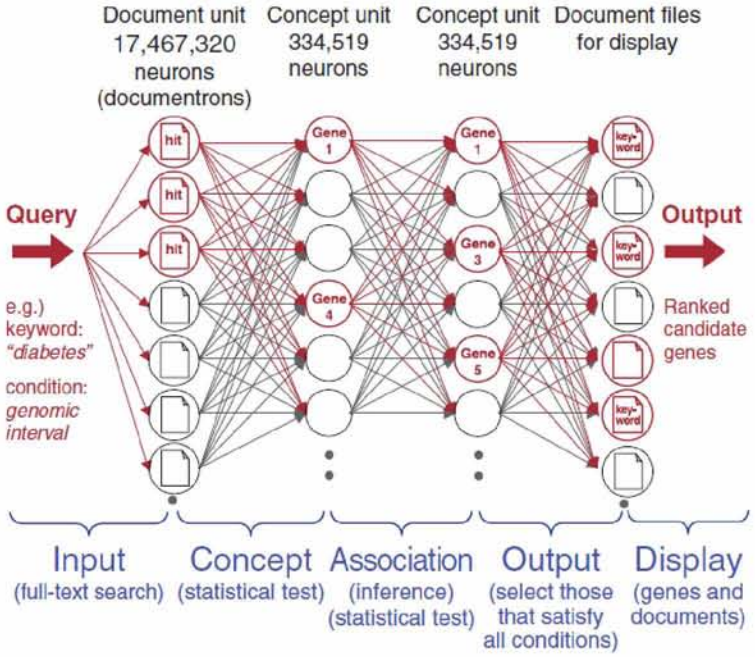
$$P\text{-}value\_Total = 1-(1-Ps)*(1-Pr)$$

όπου Ps είναι το p-value από την πρώτη συσχέτιση και Pr είναι το p-value από την δεύτερη συσχέτιση. Αυτό υπολογίζεται για να δείξει τη σημαντικότητα μεταξύ των λέξεων-κλειδιών και των γονιδίων στο πρώτο βήμα της αναζήτησης. Τέλος, ο ταξινομητής GRASE δημιουργεί μια λίστα με όλα τα ιεραρχημένα υποψήφια γονίδια με βάση το συνολικό τους p-value.

Ένα γράφημα του αλγορίθμου αναζήτησης είναι επίσης χρήσιμο για την κατανόηση της ισχύς του συστήματος. Ανάλογα με το δίκτυο των νευρώνων που σηματοδοτούνται με άλλους νευρώνες μέσω των συνδέσεων, κάθε έγγραφο θεωρείται ως νευρώνας «documentron» και τα σημαδεμένα μονοπάτια ως μια λέξη-κλειδί που ταιριάζει με το περιεχόμενο των εγγράφων (Εικόνα 1.1.6.1, Input). Τα σημαδεμένα μονοπάτια που συνδέουν κάθε νευρώνα/έγγραφο «documentron» αξιολογήθηκαν στατιστικά στο επόμενο στρώμα, υπολογίζοντας τη σημασία των ενώσεων μεταξύ των

λέξεων-κλειδιών και των γονιδίων που αναφέρονται σε σημαδεμένα έγγραφα (Εικόνα 1.1.6.1, Concept).

Εικόνα 1.1.6.1: Ενδεικτικό μοντέλο νευρωνικού δικτύου αλγόριθμου PosMed για τα υποψήφια γονίδια της ασθένειας του διαβήτη.



Μόνο οι νευρώνες (τα γονίδια) που έχουν p-value <1% (προεπιλεγμένο) παρουσιάζονται στα επόμενα σηματοδοτημένα μονοπάτια του επόμενου νευρωνικού στρώματος, ανάλογα με την δύναμη της σχέσης μεταξύ γονιδίου-γονίδιου ή των εγγράφων που μοιράζονται από κοινού (Εικόνα 1.1.6.1, Association). Οι διάφορες σχέσεις μπορεί να είναι όπως σχέσεις μεταξύ αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνης, co-expressions και ορθόλογων γονιδίων και αποτελούν τις επόμενες πρόσθετες συσχετίσεις . Μόνο τα σημαντικά γονίδια που βρίσκονται εντός μίας συγκεκριμένης περιοχής που ορίζει ο χρήστης στο γονιδίωμα εμφανίζονται μαζί με τα πλέον κατάλληλα έγγραφα που περιέχουν τα αποδεικτικά στοιχεία (Εικόνα 1, Output). Οι λέξεις-κλειδιά επισημαίνονται στα έγγραφα (Εικόνα 1, Display).

1.1.8 Αλγόριθμος Candid.

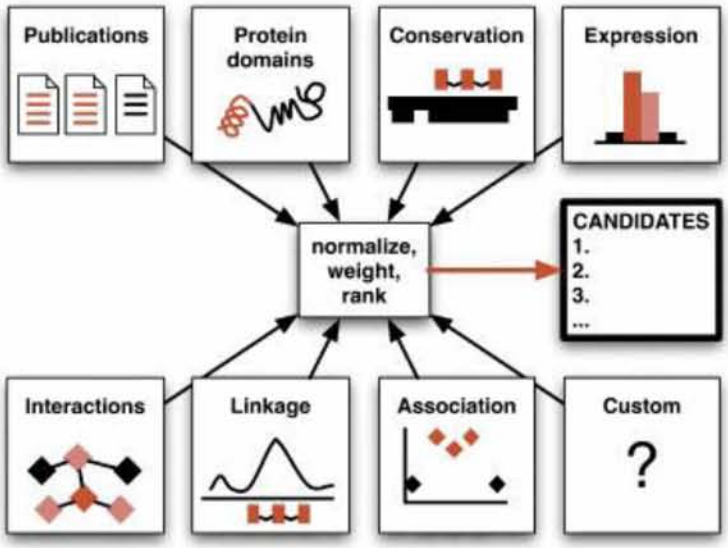
Ο αλγόριθμος Candid[111] είναι ένας ιεραρχικός αλγόριθμος που σχεδιάστηκε να παράγει αμερόληπτα αποτελέσματα και αξιόπιστες ταξινομήσεις των υποψήφιων γονιδίων που επηρεάζουν πολύπλοκα ανθρώπινα χαρακτηριστικά. Σε κάθε ανάλυση του χρησιμοποιεί πληροφορίες από δημοσιεύσεις/έρευνες, πρωτεϊνικές περιοχές, πρότυπα έκφρασης γονιδίων και δεδομένα από πρωτεϊνικές αλληλεπιδράσεις. Επιπλέον σε κάθε ανάλυση, υπάρχει η δυνατότητα να εισάγουμε και δικά μας δεδομένα ή ακόμη να ανατρέξει δεδομένα μέσα από συγκεκριμένες έρευνες. Ο αλγόριθμος Candid έχει δοκιμαστεί σε γνωστά γονίδια πολύπλοκων χαρακτηριστικών χρησιμοποιώντας δεδομένα από την βάση δεδομένων OMIM[24] και αξιολογήθηκε για το καλά διαμορφωμένο γραφικό του περιβάλλον αλλά και για τα υποψήφια ιεραρχημένα γονίδια που παρουσιάζει ανάλογα με τον συσχετισμό των χαρακτηριστικών τους. Με βάση τα αποτελέσματα που παράγει έδειξε υψηλή ευαισθησία και ειδικότητα, πράγμα που αποδεικνύει το πόσο αξιόπιστα είναι τα αποτελέσματα που μας παρουσιάζει σε σχέση με τους άλλους αλγόριθμους αναζήτησης που υπάρχουν.

Ο αλγόριθμος Candid είναι ένας αποτελεσματικός αλγόριθμος αναγνώρισης υποψήφιων γονιδίων σχεδιασμένος στον να αναλύει πολύπλοκα ανθρώπινα γενετικά χαρακτηριστικά. Για την εξεύρεση δεδομένων ανάλογα με τα γονίδια που αναζητεί κάνει αναζήτηση σε διαφορετικές πηγές δεδομένων, όπου μέσα από τον συνδυασμό των δεδομένων που εξάγουμε από αυτές τις πηγές μειώνονται κατά πολύ τα μεροληπτικά δεδομένα. Επιπλέον δίνεται δυνατότητα στον χρήστη να επιλέξει τις βάσεις δεδομένων που θα χρησιμοποιήσει ο αλγόριθμος αλλά και πόσο μεγάλη βαρύτητα θα δίνει στα δεδομένα της κάθε βάσης δεδομένων.

1.1.8.1 Μεθοδολογία αλγόριθμου Candid.

Τα αποτελέσματα της βαθμολόγησης του κάθε υποψήφιου γονιδίου παράγονται από τον συνδυασμό 8 παραμέτρων που ελέγχει για το κάθε γονίδιο ο αλγόριθμος ξεχωριστά και είναι: δημοσιεύσεις/έρευνες[39], ανάλυση του κάθε είδους με βάση την φυλογενετική ανάλυση και τα ομόλογα γονίδια του[112](cross-species conservation), πρότυπα έκφρασης γονιδίων[113], πρωτεϊνικές αλληλεπιδράσεις[7], σύνδεση αναλυμένων αποτελεσμάτων, συσχέτιση αποτελεσμάτων ανάλυσης[34] και άλλα προσαρμοσμένα δεδομένα. (Εικόνα 1.1.8.1)

Εικόνα 1.1.8.1: Παράμετροι μέσα από τους οποίους παράγεται η τελική βαθμολογία μέσα από την οποία ιεραρχείται το κάθε υποψήφιο γονίδιο.



Κάθε γονίδιο λαμβάνει διαφορετικά κριτήρια βαθμολόγησης, που ομαλοποιούνται και σταθμίζονται ανάλογα από τον χρήστη και στην συνέχεια αθροίζονται για να δώσουν στο γονίδιο την τελική του βαθμολόγηση. Τα γονίδια κατατάσσονται με βάση την τελική βαθμολογία που συγκεντρώνει το καθένα και παρουσιάζονται σε μία λίστα όλα μαζί ιεραρχημένα από το πιο υψηλά συσχετισμένο με μία ασθένεια μέχρι αυτό που έχει την μικρότερη πιθανότητα να σχετίζεται με μια ασθένεια. Ενώ ο αλγόριθμος Candid αξιολογεί όλα τα ανθρώπινα γονίδια από προεπιλογή, δίνεται η δυνατότητα στους χρήστες να περιορίσουν την ανάλυση τους μόνο με πρωτεϊνικά κωδικοποιημένα γονίδια.

1.1.8.2 Βαθμολόγηση δεδομένων μέσα από τις δημοσιεύσεις.

Οι δημοσιευμένες βάσεις δεδομένων του αλγόριθμου Candid αποτελούνται από τις άμεσες συνδέσεις μεταξύ PubMed IDs[39] και EntrezGene Ids. Ένα γονίδιο συνδέεται με μία δημοσίευση όταν η δημοσίευση περιγράφει στοιχεία της ακολουθία ή της λειτουργικότητας του. Οι δημοσιεύσεις που σχετίζονται με το X γονίδιο προσδιορίζονται ως το σύνολο G(x). Ο χρήστης παρέχει ένα σύνολο απο λέξεις-κλειδιά που σχετίζονται με το χαρακτηριστικό ή την ασθένεια που τον ενδιαφέρει, που θα είναι κατάλληλα για μια τυπική αναζήτηση βιβλιογραφίας. Οι λέξεις-κλειδιά μπορούν να κυμαίνονται από πολύ χαρακτηριστικές μέχρι και πολύ ειδικές για να πιο σωστές.

Αναμένεται ότι ορισμένα χαρακτηριστικά θα έχουν μόνο λίγες σχετικές λέξεις-κλειδιά, και σε αυτές τις περιπτώσεις, προσθέτοντας περισσότερες λέξεις-κλειδιά, που είναι περισσότερο αόριστες στην συσχέτιση, τότε θα εισάγουμε πιθανόν περισσότερο θόρυβο στην τελική κατάταξη των υποψήφιων γονιδίων. Επίσης, τα συνώνυμα θα πρέπει να συμπεριλαμβάνονται όταν χρειάζεται να συμπεριλάβουμε όσο το δυνατό περισσότερες σχετικές δημοσιεύσεις.Το αποτελεσματικότερο σύνολο λέξεων-κλειδιών χρησιμοποιείται σε ένα «κείμενο – λέξη» στην αναζήτηση του PubMed για να προσδιορίσει όλες τις αντιστοιχίσεις στις δημοσιεύσεις (M). Η βαθμολόγηση της δημοσίευσης P, για το γονίδιο X κυμαίνετε μεταξύ 0 και 1 και δίνεται από τον ακόλουθο τύπο:

$$P(X) = \frac{|M \cap G(X)|}{|G(X)|}$$

Το σκεπτικό για τη μέθοδο αυτή είναι να ανταμείψει τα υποψήφια γονίδια που συνήθως συνδέονται με τις δημοσιεύσεις που περιγράφουν το χαρακτηριστικό που μας ενδιαφέρει , ανεξάρτητα από το βαθμό στον οποίο τα εν λόγω γονίδια χαρακτηρίζονται στη βιβλιογραφία. Μετά όλα τα υποψήφια γονίδια βαθμολογούνται για αυτό το κριτήριο και η βαθμολογία τους κανονικοποιήται από την διαίρεση τους.

1.1.8.3 Βαθμολόγηση πρωτεϊνικής περιοχής

Οι πληροφορίες της πρωτεϊνικής περιοχής λαμβάνονται από την NCBI[37] βάση δεδομένων και τη βάση πρωτεϊνικών περιοχών CDD[114], μια επιμεληημένη βάση δεδομένων που ενσωματώνει στοιχεία από άλλες βάσεις δεδομένων, όπως Pfam[115], SMART[116] και COG[117]. Η CDD βάση δεδομένων είναι παρόμοια με την InterPro βάση δεδομένων. Οι καταχωρήσεις με βάση τις πρωτεϊνικές περιοχές διαθέτουν περιγραφή της πρωτεϊνικής περιοχής και αλλά και του τρόπου συνδέσεις με τα γονίδια τα οποία περιέχουν τις αλληλουχίες αμινοξέων που αποτελούν αυτούς τις πρωτεϊνικές περιοχές. Όπως και σε άλλες δημοσιεύσεις πρέπει να τηρούμε κάποιες προϋποθέσεις, δηλαδή ο χρήστης παρέχει ένα σύνολο από λέξεις-κλειδιά που αφορούν τα χαρακτηριστικά, τα οποία χρησιμοποιούνται για αναζήτηση κάποιας περιγραφής πρωτεϊνικής περιοχής. Κάθε γονίδιο του οποίου η μεταφρασμένη αλυσίδα με αμινοξέα, περιέχει και ταιριάζει με μία τουλάχιστον πρωτεϊνική περιοχή που αναζητούμε τότε λαμβάνει την βαθμολογία 1. Όλα τα άλλα γονίδια λαμβάνουν την βαθμολογία 0.

Η μέθοδος αυτή χρησιμεύει για να προβλέψουμε τις πιθανολογούμενες λειτουργίες των γονιδίων, και δεδομένου ότι η πρόβλεψη της πρωτεϊνικής περιοχής

γίνετε με βάση την ακολουθία των βάσεων, έχουν τη δυνατότητα να προβλέψουν σχεδόν όλα τα πρωτεϊνικά κωδικοποιημένα γονίδια για τα οποία οι πρωτεϊνικές περιοχές έχουν εντοπιστεί, ανεξάρτητα από το πόσο εκτεταμένα θα έχουν χαρακτηριστεί στην επιστημονική βιβλιογραφία.

1.1.8.4 Βαθμολόγηση παραμέτρων ομόλογων και φυλογενετικής ανάλυσης.

Η βάση δεδομένων HomoloGene της NCBI[37] παρέχει τις πληροφορίες για τις παραμέτρους ομόλογων και φυλογενετικότητας που χρησιμοποιούνται για την ανάλυση που κάνει ο αλγόριθμος CANDID. Η HomoloGene βάση δεδομένων αναλύει τα γονίδια από 18 ολοκληρωμένους σε δεδομένα αλληλουχίας οργανισμούς και ανιχνεύει τα ομόλογά χρησιμοποιώντας τα αμινοξέα και τις αλληλουχίες DNA. Για κάθε γονίδιο, ένα από τα εννέα πεδία παρέχεται για να περιγράψει την φυλογενετική ομοιότητα μεταξύ των ανθρώπων και των οργανισμών βρίσκοντας και τα πιο πιθανά ομόλογα τους.

Τα ειδικά ανθρώπινα γονίδια λαμβάνουν βαθμολογία 0, ενώ τα γονίδια με την σήμανση «Eukaryota» λάβουν τη μέγιστη βαθμολογία 1. Βαθμολογίες για τα υπόλοιπα επτά πεδία τύπων γονιδίων κατανέμονται ομοιόμορφα, με βαθμολογίες που κυμαίνονται από 0,125 με 0,875. Η υψηλή βαθμολογία για αυτό το κριτήριο μπορεί να έχει σημασία με το φαινότυπο για τον οποίο είναι γνωστό ότι από αυτόν προβλέπουμε κυτταρικές διαδικασίες, όπως ο καρκίνος, ενώ μπορεί να είναι άσχετο με το φαινότυπο όπως χρώμα μαλλιών που επηρεάζει μόνο ένα μικρό υποσύνολο των ειδών. Οι Βαθμολογίες για αυτό το κριτήριο κανονικοποιούνται από την διαίρεση όλων των βαθμολογίων από τις παραμέτρους συντήρησης των γονιδίων με την υψηλότερη βαθμολογία διατήρησης.

1.1.8.5 Βαθμολόγηση πρότυπων έκφρασης γονιδίων.

Ο αλγόριθμος Candid εκτελεί τις διάφορες αναζητήσεις του για να μαζέψει πληροφορίες για το πρότυπο έκφρασης του κάθε γονιδίου από την βάση δεδομένων GNF, που περιλαμβάνει τα επίπεδα έκφρασης των 17.761 ανθρώπινων γονιδίων σε 79 ανθρώπινους ιστούς. Συγκρίνει κάθε επίπεδο γονιδιακής έκφρασης σε ολόκληρη την βάση δεδομένων. Κάθε γονίδιο λαμβάνει την βαθμολογία 1 για τους ιστούς που περιλαμβάνονται στο γονίδιο και εκφράζονται πιο πολύ, ενώ οι βαθμολογίες των γονιδίων για τους υπόλοιπους ιστούς αντιστοιχούν στο λόγο του επιπέδου έκφρασης σε αυτό τον ιστό με το ανώτατο επίπεδο έκφρασης του γονιδίου.

Τα γονίδια που εκφράζονται ειδικά σε ένα συγκεκριμένο ιστό λαμβάνουν μια υψηλή βαθμολογία για το εν λόγο ιστό και πολύ χαμηλές βαθμολογίες για όλους τους άλλους ιστούς, ενώ τα γονίδια που έχουν περίπου ίσα επίπεδα έκφρασης σε όλους τους ιστούς λαμβάνουν περίπου τη ίδια βαθμολογία σε όλους τους ιστούς. Η μέθοδος αυτή χρησιμεύει για να τονίσει τα γονίδια με ιστούς που πιστεύεται ότι παίζουν σημαντικότερο ρόλο. Η βαθμολογία που συγκεντρώνει το κάθε γονίδιο με βάση το πρότυπο έκφρασης αντιστοιχεί στο άθροισμα της βαθμολογίας των ιστών που βρέθηκε να ταιριάζει. Οι Βαθμολογίες για αυτό το κριτήριο κανονικοποιούνται από την διαίρεση όλων βαθμολογιών έκφρασης του γονιδίου με την υψηλότερη βαθμολογία έκφρασης .

1.1.8.6 Βαθμολόγηση με βάση τις αλληλεπιδράσεις των πρωτεϊνών.

Οι αλληλεπιδράσεις μεταξύ πρωτεϊνών παρέχουν πολύτιμες πληροφορίες για την λειτουργικότητα ενός γονιδίου, ακόμη και αν δεν υπάρχουν άλλες πληροφορίες. Η πληροφορίες σχετικά με τις αλληλεπιδράσεις των πρωτεϊνών προέρχονται κατά κύριο λόγο από την NCBI[37] βάση δεδομένων αλλά και από τις βάσεις δεδομένων BIND[12], HPRD[62] και BioGRID[118]. Για τον προσδιορισμό του βαθμού αλληλεπιδράσεις ενός γονιδίου, ο αλγόριθμος CANDID αθροίζει τις βαθμολογίες των δημοσιεύσεων και πρωτεϊνικών περιοχών για όλα τα γονίδια που έχουν αλληλεπιδράσεις.

1.1.8.7 Βαθμολόγηση δεδομένων σύνδεσης.

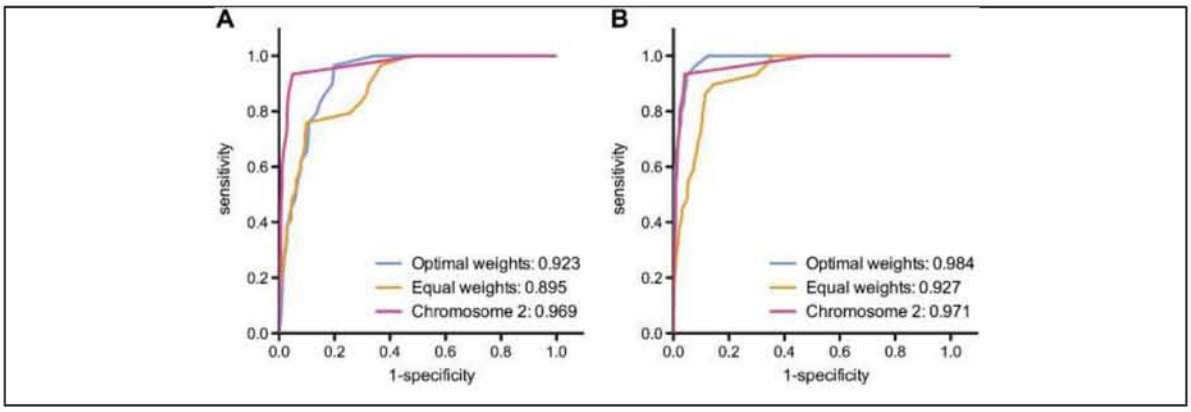
Ο αλγόριθμος CANDID είναι ειδικά σχεδιασμένος για να δέχεται ένα ή περισσότερα σύνολα από τα υφιστάμενα δεδομένα σύνδεσης. Με βάση τις πληροφορίες σύνδεσης μπορούμε να περιορίσουμε τις αναλύσεις των γονιδίων που θα παρουσιάσει ο αλγόριθμος CANDID σε μία ή περισσότερες περιοχές ή να δώσουμε προτεραιότητα σε γονίδια, που είναι βασισμένα στην πλησιέστερη LOD βαθμολογία (Απόσταση μεταξύ δύο περιοχών loci που βρίσκονται στο ίδιο χρωμόσωμα) του κάθε γονιδίου. Με τον αλγόριθμο αυτό εύκολα μπορούμε να συνδέσουμε αρχεία στην ανάλυση μας, από τους αλγόριθμους GENEHUNTER, MERLIN και SOLAR. Κατά την διαδικασία σύνδεσης δημιουργείται ένα αρχείο αθροίζοντας τις επιμέρους καμπύλες LOD. Όλες οι σημαδεμένες θέσεις είναι σε centiMorgans (cm) και αντιστοιχούν στον γενετικό χάρτη Marshfield. Ο αλγόριθμος CANDID καθορίζει την βαθμολογία σύνδεσης για το κάθε γονίδιο που βασίζεται στην προσεγγιστική LOD βαθμολογία από την θέση του γονιδίου.

Οι προσεγγιστικές cM (centiMorgans)θέσεις για κάθε γονίδιο παρεμβάλλονται εξετάζοντας την φυσική απόσταση μεταξύ του μέσου γονιδίου και των δύο πλησιέστερων Marshfield δεικτών. Η βαθμολογία σύνδεσης ενός γονιδίου παρεμβάλλεται μεταξύ των δύο πιο κοντινών δεικτών με την καταχωρημένη LOD βαθμολογία τους και στη συνέχεια ομαλοποιούνται διαιρώντας το με την υψηλότερη ατομική βαθμολογία σύνδεσης που πήρε κάποιο γονίδιο. Η υψηλότερη βαθμολογία σύνδεσης είναι 1, αλλά μερικά γονίδια μπορεί να έχουν και αρνητική βαθμολόγηση σύνδεσης εφόσον σχετίζονται με αρνητικές LOD βαθμολογίες. Αυτή η μέθοδος επομένως φιλτράρει έξω κάποια γονίδια, τα οποία μπορεί να παρουσίασαν μέτριες έως υψηλές βαθμολογίες σε άλλα κριτήρια.

1.1.8.8 Βαθμολόγηση δεδομένων συσχέτισης.

Ο αλγόριθμος CANDID χρησιμοποιήσει μονό δεδομένα συσχετισμού που βασίζονται σε νουκλεοτιδικούς πολυμορφισμούς (SNP). Ο χρήστης παρέχει ένα αρχείο όπου κάθε γραμμή περιέχει το χαρακτηριστικό ID του κάθε SNP που έχει με βάση την SNP βάση δεδομένων και το p-value(τιμή) που δίνεται για αυτό το SNP. Ο αλγόριθμος CANDID αναθέτει το κάθε ένα από αυτά τα SNPs στα συσχετισμένα γονίδιά τους, σύμφωνα με την βάση δεδομένων SNP που έχει η NCBI βάση δεδομένων. Η καλύτερη τιμή(p-value) για κάθε γονίδιο παρακρατάτε από 1 έως τη παραγόμενη για κάθε γονίδιο βαθμολογία συνάφειας. Επίσης σημειώνεται το γονίδιο με ένα SNP και πολύ ασήμαντη τιμή(p-value) και λαμβάνει πάντα μια υψηλότερη βαθμολογία από ένα γονίδιο που δεν σχετίζεται με κάποιο SNPs.

Εικόνα 1.1.8.2: Επιτυχής προβλέψεις σε γονίδια της βάσης δεδομένων OMIM που χρησιμοποιούν πέντε βασικές παράμετρους. Παράγονται οι καμπύλες ROC για (A) όλο το γονιδίωμα και (B) αναλύσει συγκεκριμένου χρωμοσώματος.



1.1.9 Αλγόριθμος GenaCards.

Ο αλγόριθμος GeneCards[119, 120] είναι μία διαδικτυακή εφαρμογή εξόρυξης ανθρώπινων γονιδίων και κωδικοποιημένων πρωτεϊνών, με βάση λειτουργιών της γονιδιωματικής και ιατρικών- βιολογικών ερευνών, που σχετίζονται με ασθένειες. Ο αλγόριθμος GeneCards παρέχει συνοπτικές πληροφορίες για τη δομή και την λειτουργία των ανθρώπινων γονιδίων. Εξάγει και ενσωματώνει ένα επιλεγμένο υποσύνολο από πληροφορίες σε κάθε γονίδιο που αναζητά, το οποίο λαμβάνεται από διάφορες βάσεις δεδομένων και δημοσιευμένες έρευνες.

Τα διάφορα πεδία που κάνει τις αναζητήσεις του για κάθε γονίδιο χωρίζονται σε αναζητήσεις με βάση (i) τις λέξεις-κλειδία – συνώνυμα που υπάρχουν, (ii) τα γονίδια που βρέθηκε να σχετίζονται με τις λέξεις-κλειδία που εισάγαμε κάνει μια έρευνα για τις χρωμοσωμικές περιοχές που βρίσκονται και με βάση αυτές κάνει περεταίρω αναζητήσεις, (iii) τα γονίδια που βρέθηκαν και τις οντολογίες που φέρουν εκτελεί αναζητήσεις με βάση τις οντολογίες, (iv) τα γονίδια που βρέθηκαν και τις εκφράσεις που φέρουν εκτελεί αναζητήσεις με βάση τις εκφράσεις γονιδίων και τέλος (v) τα γονίδια που βρέθηκαν εκτελεί αναζητήσεις για τους μονούς νουκλεοτιδικούς πολυμορφισμούς και τους μεταξύ τους συσχετισμούς.

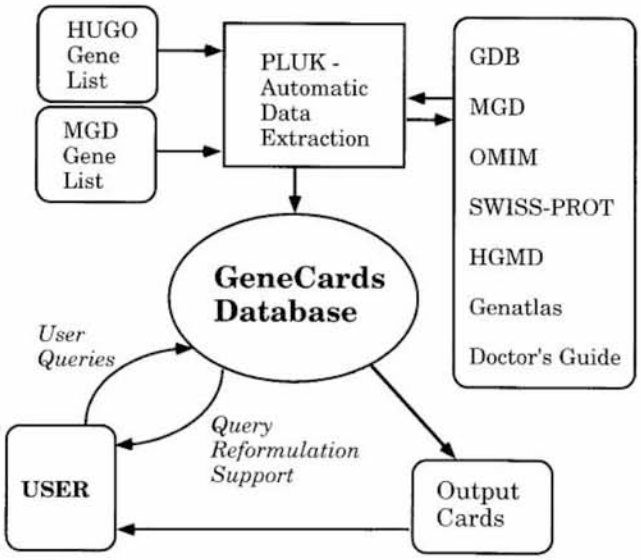
1.1.9.1 Μεθοδολογία αλγόριθμου.

Με τις λέξεις κλειδιά που εισάγονται στον αλγόριθμο GeneCards, γίνονται διάφορες αναζητήσεις στις βάσεις δεδομένων NCBI και PubMed και μαζεύονται οι σχετικές πληροφορίες. Αφού μαζέψει όλες τις σχετικές πληροφορίες με την χρήση έξυπνων μεθόδων εξόρυξης δεδομένων, τότε εκτελεί αναζητήσεις στις βάσεις δεδομένων HUGO και GDB όπου μαζεύει όλα τα σχετικά γονίδια που βρέθηκε να σχετίζονται με τις πληροφορίες από τις προηγούμενες βάσεις δεδομένων με βάση τις λέξεις κλειδιά. (Εικόνα 1.1.9.1)

Αφού δημιουργηθεί μία πλήρης λίστα με όλα τα γονίδια που βρέθηκαν να σχετίζονται με τις λέξεις κλειδιά, τότε ο αλγόριθμος εκτελεί αναζητήσεις στις βάσεις δεδομένων SWISS-PROT, OMIM, Genatlas , GDB, HGNC , NCBI , ENSEMBL και UniProtKB μέσα από τις οποίες συλλέγει πληροφορίες για το κάθε γονίδιο με βάση τα

πεδία ενδιαφέροντος που χωρίζεται ο αλγόριθμος(χρωμοσωμικές περιοχές, οντολογίες, πρότυπα έκφρασης γονιδίων, νουκλεοτιδικούς πολυμορφισμούς) (Εικόνα 1.1.9.1)

Εικόνα 1.1.9.1: Αναπαράσταση του τρόπου εξαγωγής γονιδίων και των σχετικών πληροφοριών τους που σχετίζονται με τις λέξεις κλειδιά που εκτελείται στον αλγόριθμος GeneCards.



1.1.9.2 Αναζήτηση δεδομένων στον αλγόριθμο GenCards.

Γενικά οι αναζητήσεις στον αλγόριθμο Gene Cards εκτελούνται πολύ γρήγορα. Σε κάθε αναζήτηση που εκτελούμε με τις λέξεις κλειδιά , αναζητούνται πληροφορίες κατά μήκος πολλαπλών διαδρομών και τα αποτελέσματα παράγονται με την χρήση πολλών διαθέσιμων βάσεων δεδομένων.

Ένα μεγάλο μειονέκτημα του αλγόριθμου GeneCards είναι ότι δεν εκτελεί καμία μεθοδολογία για να ιεραρχήσει τα γονίδια που βρίσκει να σχετίζονται με μία ασθένεια , αλλά απλά τα παρουσιάζει ανάλογα με το πόσο πολύ ανιχνεύτηκε το κάθε γονίδιο στις διάφορες βάσεις δεδομένων που αναζητεί. Δηλαδή παρουσιάζεται πρώτο το γονίδιο που βρέθηκε να είναι καταχωρημένο σε όλες τις βάσεις δεδομένων και ανάλογα παίρνουν σειρά τα υπόλοιπα γονίδια.

Για τον λόγο ότι συχνά, η λέξη κλειδί που αναζητούμε βρίσκεται σε δεκάδες χώρους, καθιστάτε αδύνατο, στις περισσότερες περιπτώσεις, να προσδιοριστεί από πια συγκεκριμένα πεδία αντλήθηκε η πληροφορία και η σχετική βαθμολογία που δίνεται από την βάση δεδομένων που προέρχεται. Για αυτό το λόγο ο αλγόριθμος αναζήτησης GeneCards παρέχει δύο φάσεις αναζητήσεις, και ενεργοποιούνται με δύο συγκριτικά

επίπεδα (δείκτες). Δηλαδή το όλο πρόβλημα επιλύνεται κάνοντας χρήση δύο δεικτών, το κύριο δείκτη και το δευτεροβάθμιο δείκτη.

- Στον πρώτο και πιο πάνω σε επίπεδο δείκτη περιέχονται όλες οι σχετικές πληροφορίες που βρίσκουμε για τις λέξεις κλειδιά(για παράδειγμα ονοματολογίες γονιδίων).

-Στον δεύτερο δείκτη πιο κάτω επίπεδο παρέχονται οι σχετικές πληροφορίες που βρέθηκαν για τις πληροφορίες που φέρει ο πρώτος δείκτης (για παράδειγμα φέρει πληροφορίες που έχουν να κάνουν με το κάθε γονίδιο που βρίσκεται στον πρώτο δείκτη).

Σε μία τυπική αναζήτηση που εκτελείται στον αλγόριθμο GeneCards για γονίδια που σχετίζονται με ασθένειες εκτελούνται τα εξής βήματα:

(i) Στον πρώτο δείκτη αποθηκεύονται όλα τα ονόματα των γονιδίων που βρέθηκε να περιέχουν μέσα την λέξη κλειδί που εισάγαμε.

(ii) Όταν ο χρήστης ζητά να ανοίξει και να μελετήσει περεταίρω τα δεδομένα και τις σχετικές πληροφορίες που βρέθηκαν για ένα από τα γονίδια που παρουσιάζονται στην λίστα του πρώτου δείκτη, παρουσιάζεται μια λίστα με λεπτομερή δεδομένα και τις βάσεις δεδομένων από όπου προέρχονται.

1.2 Σύγκριση Μεθόδων / Αλγορίθμων.

Στην παρούσα εργασία χρησιμοποιήσαμε τους αλγόριθμους ιεραρχικής αναζήτησης υποψηφίων γονιδίων που σχετίζονται με ασθένειες Prospectr, Suspects, Gene-Prospectr, Phenopred, Posmed, Candid, SNPs3d, FITsSNPs και GeneCards. (Πίνακας 1.2.1) Ο λόγος για τον οποίον επιλέχτηκαν οι πιο πάνω αλγόριθμοι για την διεκπεραίωση της εργασίας μας ήταν γιατί και στους εννέα αυτούς αλγόριθμους μπορούσαμε να κάνουμε απλές αναζητήσεις για υποψήφια γονίδια που σχετίζονται με μία ασθένεια απλά εισάγοντας όλες τις λέξεις-κλειδιά που σχετίζονταν με την ασθένεια που θέλαμε να αναζητήσουμε.

Πίνακας 1.2.1: Ιεραρχικοί αλγόριθμοι αναζήτησης που χρησιμοποιήθηκαν στην εργασία μας.

Αλγόριθμος	Σύνδεσμος
Prospectr	http://www.genetics.med.ed.ac.uk/prospectr/
GeneProspector	www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do
Suspects	www.genetics.med.ed.ac.uk/suspects/search.shtml
Phenopred	www.phenopred.org
PosMed	http://omicspace.riken.jp/PosMed/search
Candid	http://dsgweb.wustl.edu/hutz/index.html
SNP3D	http://www.SNPs3D.org
FitSNPs	http://fitsnps.stanford.edu/
Genecards	http://www.genecards.org

Όλοι οι αλγόριθμοι μας παράγουν μία λίστα με όλα τα ιεραρχημένα γονίδια. Το μέγεθος τις κάθε λίστας ως προς τον αριθμό των γονιδίων που συμπεριλαμβάνονταν μέσα σε αυτή που μας επέστρεφε ο κάθε αλγόριθμος δεν ήταν πάντα το ίδιο, μερικές φορές κυμαινόταν από μερικές δεκάδες και άλλες φορές κυμαινόταν σε εκατοντάδες γονίδια. Αυτό φυσικά οφείλεται στο πόσο μεγάλο εύρος βάσεων δεδομένων και γονιδίων αναζητούσε ο κάθε αλγόριθμος ξεχωριστά, αλλά και στις παραμέτρους μέσα από τις οποίες συσχέτιζε τα γονίδια του με την ασθένεια που αναζητούσαμε. Στον πίνακα 1.2.2 παρουσιάζονται οι αλγόριθμοι αναζήτησης μαζί με τις βάσης δεδομένων που χρησιμοποιεί ο καθένας από αυτός ξεχωριστά. Είναι αξιοσημείωτο ότι και οι 9 αλγόριθμους που χρησιμοποιούμε αναζητούν δεδομένα για τα υποψήφια γονίδια που σχετίζονται με ασθένειες από την OMIM βάση δεδομένων.

Πίνακας 1.2.2: Βάσεις δεδομένων που χρησιμοποιούνται από τους εννέα διαφορετικούς αλγόριθμους.

Α/Α	ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ	ΑΛΓΟΡΙΘΜΟΙ ΠΟΥ ΤΗΝ ΧΡΗΣΙΜΟΠΟΙΟΥΝ
1	OMIM	GeneProspector, Suspects, Prospectr, SNP3Ds, FITsSNP, PosMed, Phenopred, Candid, GeneCard
2	GO	Suspects, Prospectr, Phenopred, SNP3Ds, FITsSNP, Phenopred, PosMed,
3	Enter Gene	GeneProspector, Candid
4	Ensembl Human	GeneProspector, Prospectr,
5	Swiss-Prot	Phenopred, Prospectr
6	HPRD	Phenopred, Candid
7	HUGE Navigator	GeneProspector,
8	NCBI	GeneProspector, Suspects, Prospectr, SNP3Ds, FITsSNP,
9	HGMD	GeneProspector, Suspects, Prospectr, SNP3Ds, FITsSNP,
10	HomoloGene	Prospectr, Candid,
11	PubMed	GeneProspector, SNP3Ds, Candid, PosMed
12	SNP	GeneProspector, SNP3Ds, FITsSNP, Candid
13	GAD	GeneProspector, Suspects, FITsSNP,
14	GWAS	GeneProspector, FITsSNP
15	InterPro	Prospectr, Candid
16	BIND	SNP3Ds, Candid
17	PFam	GeneProspector, Candid
18	BioGrid	GeneProspector, Candid
19	DO	Phenopred
20	PROSITE	Phenopred
21	OPHID	Phenopred
22	NGEA	Prospectr
23	RefSeg	SNP3Ds
24	CDD, SMART	Candid
25	GNF	Candid

Σημαντικό ρόλο για την πρόγνωση και ιεράρχηση των υποψηφίων γονιδίων παίζει ο τρόπος μέσα από τον οποίο αναζητούνται αλλά και βαθμολογούνται τα υποψήφια γονίδια. Μετά από διερεύνηση του τρόπου λειτουργίας του κάθε αλγόριθμου ανακαλύψαμε ότι ο κάθε αλγόριθμος εξάγει τα δικά του αποτελέσματα από τον έλεγχο διαφορετικών βιολογικών παραμέτρων κάθε φορά. Αυτό είναι πολύ σημαντικό κομμάτι

τις εργασίας μας, γιατί έτσι δεν θα έχουμε μονόπλευρα αποτελέσματα, αλλά θα έχουμε αποτελέσματα που θα συνδυάζονται από πάρα πολλές διαφορετικές βιολογικές παραμέτρους αλλά και μεθοδολογίες, και στην περίπτωση που ένα γονίδιο εμφανίζεται σε όλους τους αλγόριθμους αναζήτησης, πέρα από το ότι εμφανίζεται συχνά, θα είμαστε πλέον σίγουροι γιατί θα πληρείται μέσα από όλες τις συνθήκες ελέγχου.

Στον Πίνακα 1.2.3 παρουσιάζονται όλες οι βιολογικές παράμετροι που βρέθηκε να χρησιμοποιούν οι εννέα αλγόριθμοι αναζήτησης μας, και στον Πίνακα 1.2.4 παρουσιάζονται οι σχετικές μεθοδολογίες που χρησιμοποιούν για εύρεση, ανάλυση εκπαίδευση και ιεράρχηση των υποψηφίων γονιδίων που μας παρουσιάζουν.

Πίνακας 1.2.3: Βιολογικές παράμετροι που εξετάζει ο κάθε αλγόριθμος αναζήτησης που χρησιμοποιούμε στην εργασία μας.

Α/Α	Παράμετροι Ελέγχου	Αλγόριθμος
1	Δημοσιευμένη Βιβλιογραφία	GeneProspector, Candid, PosMed, SNP3Ds,GeneCards
2	Μοριακό & Λειτουργικό επίπεδο πρωτεϊνών	PosMed, Phenopred, SNP3Ds, GeneCards
3	Κωδικοποιημένη ακολουθία πρωτεϊνών	Phenopred, Prospectr, Suspects, PosMed, Candid, GeneCards
4	Οντολογίες γονιδίων	Prospectr, Suspects, Phenopred
5	Πρότυπα έκφρασης γονιδίων (Μικροσυστοιχίες)	PosMed, Suspects, Candid, GeneProspector, SNP3Ds, FITsSNP
6	Νουκλεοτιδικοί πολυμορφισμοί	GeneProspector, SNP3Ds, FITsSNP, Prospectr
7	Πρωτεϊνική περιοχή	Candid, Prospectr, Suspects
8	Φυλογενετικότητα, Παράλογα, Ομόλογα	Prospectr, Suspects, FITsSNP, Candid
9	Φαινοτυπικές παραλλαγές	Suspects, PosMed
10	Πειραματικά δεδομένα από ζώα	PosMed, GeneProspector, Prospectr
11	Λειτουργικός σχολιασμός γονιδίων	Prospectr, Suspects, GeneCards
12	Χαρακτηριστικά ακολουθίας γονιδίου	Prospectr, Suspects, SNP3Ds, GeneCards
13	Αριθμός εξονίων	Prospectr, Suspects,
14	Δίκτυα αλληλεπίδρασης πρωτεϊνών PPIs	Phenopred, SNP3Ds
15	Ανάλυση μικροσυστοιχιών	FITsSNP

Μέσα από την εξερεύνηση και την κατανόηση της λειτουργίας του κάθε αλγόριθμου ξεχωριστά, ανακαλύψαμε ότι κάποιοι από τους αλγόριθμους που

χρησιμοποιήσαμε συνεργάζονται με κάποιους άλλους. Για παράδειγμα ο αλγόριθμος Suspects έχει μια πολύ στενή συνεργασία με τον αλγόριθμο Prospectr. Με βάση τις λέξεις κλειδιά που δίνονται στον αλγόριθμο Suspects, ο ίδιος εκτελεί αναζήτηση στον αλγόριθμο Prospectr και φορτώνει τα δεδομένα που του επιστρέφει με τα υποψήφια γονίδια στην δική του λίστα και μετά τα ξανά σκοράρει και ιεραρχεί με βάση τις δικές του παραμέτρους. Άλλοι αλγόριθμοι που βρέθηκε να αναζητούν δεδομένα από τρίτους αλγόριθμους είναι ο GeneProspector που αναζητεί υποψήφια γονίδια κάνοντας χρήση τους αλγόριθμους SNP3Ds και GeneCards και ο αλγόριθμος Candid που αναζητεί υποψήφια γονίδια από τον Suspects.

Πίνακας 1.2.4: Σχετικές μεθοδολογίες / τεχνικές που χρησιμοποιούνται από τους αλγόριθμους εύρεσης και ιεράρχησης υποψηφίων γονιδίων.

Α/Α	**Μέθοδος Εύρεσης / Ανάλυσης υποψήφιων γονιδίων**	**Αλγόριθμος που την χρησιμοποίει**
1	Διανυσματικές μηχανές υποστήριξης (SVM)	Prospectr, SNP3Ds, FITsSNP
2	Νευρωνικά δίκτυα	PosMed
3	Δένδρα απόφασης	Prospectr, Suspects
4	Δενδρικές αναζητήσεις	GeneProspector
5	Μπεϊσιανά δίκτυα	Prospectr, Suspects
6	Έλεγχος Man Whitney U, t-test	Prospectr, Suspects, FITsSNP
7	Δίκτυα συσχετισμού γονιδίων (Y2H,KnowledgeNet)	SNP3Ds
8	Μετασχηματισμοί z-score	Phenopred

1.3. Πακέτο στατιστικής ανάλυσης RankProd.

Το πακέτο RankProd[1, 121] είναι μία βιβλιοθήκη του προγράμματος R-Project[122] όπου μπορεί να αναζητήσει εύκολα κανείς και να κατεβάσει δωρεάν από το διαδίκτυο. Περιλαμβάνει συναρτήσεις για την ανάλυση της έκφρασης γονιδίων από microarray δεδομένα, προπάντων για τον προσδιορισμό διαφορετικά εκφρασμένων γονιδίων αλλά και συσχέτιση της ιεράρχησης των γονιδίων από διάφορες πηγές και εξαγωγή των πιο συσχετισμένων γονιδίων με βάση τα βάρη τους.

Η μέθοδος RankProd χρησιμοποιεί το rank-product , μια μη παραμετρική μέθοδο για τον εντοπισμό των υπό-εκφρασμένων ή υπέρ-εκφρασμένων (up-regulated or down-regulated) γονιδίων που είναι ιεραρχημένα με βάση μία προϋπόθεση ενάντια σε μια άλλη προϋπόθεση. (π.χ. δύο διαφορετικές θεραπείες, δύο διαφορετικούς τύπους ιστών, κλπ.) Η Rank-Product είναι μια μη-παραμετρική στατιστική που ανιχνεύει τα στοιχεία που βρίσκονται σε σταθερά υψηλή θέση σε κάποιες συγκεκριμένες λίστες, για παράδειγμα τα γονίδια που βρίσκονται μεταξύ των πιο έντονα υπερεκφραζόμενων γονιδίων και είναι σταθερά σε διάφορα διεξαγόμενα πειράματα. Βασίζεται στην υπόθεση ότι, σύμφωνα με τη μηδενική υπόθεση ο βαθμός όλων των στοιχείων είναι η τυχαία πιθανότητα να βρεθεί ένα συγκεκριμένο σημείο μεταξύ της κορυφής r από τα n στοιχεία σε μια λίστα και υπολογίζεται με. $p = r/n$.Πολλαπλασιάζοντας τις πιθανότητες αυτές οδηγεί στον καθορισμό του RP που δίνεται από τον τύπο:

$$RP = \prod i \frac{r_i}{n_i}$$

οπου ri είναι ο βαθμός ιεράρχησης του στοιχείου στην i-οστή λίστα και ni είναι ο συνολικός αριθμός των στοιχείων στη ι-οστή λίστα. Όσο μικρότερη είναι η αξία του RP, τόσο μικρότερη είναι η πιθανότητα ότι η παρατηρούμενη τοποθέτηση του αντικειμένου στην κορυφή των καταλόγων οφείλεται στην τύχη. Το RankProduct είναι ισοδύναμο με τον υπολογισμό του γεωμετρικού μέσου βαθμού, αντικαθιστώντας το προϊόν με την άθροιση που οδηγεί σε στατιστικά στοιχεία (μέσος όρος βαθμού) που είναι λίγο πιο ευαίσθητος στα εξερχόμενα δεδομένα και βάζει ένα υψηλότερο ασφάλιστρο για τη συνοχή μεταξύ τις ιεράρχησης σε διάφορες λίστες.

Ο κατάλογος των υπέρ-εκφρασμένων ή υπό-εκφρασμένων γονιδίων επιλέγονται με βάση το εκτιμώμενο ποσοστό ψευδώς θετικών προβλέψεων (PfP), το οποίο είναι επίσης γνωστό ως ψευδή ρυθμός ανακάλυψης (FDR). Η μέθοδος αυτή είναι σε θέση να

αναλύσει και Affymetrix δεδομένα Genechip[123] καθώς και με cDNA δεδομένα πίνακα μετά από την κανονικοποίηση. Μια άλλη εφαρμογή αυτής της μεθόδου είναι η ικανότητά της να συνδυάζει σύνολα δεδομένων από διαφορετικές πηγές σε μία ανάλυση ώστε να αυξηθεί η ισχύς της ταυτότητας. Στην πράξη, αυτό καθιστάτε δυνατό για σύνολα δεδομένων που παράγονται σε διαφορετικά εργαστήρια ή σε διαφορετικά περιβάλλοντα να συνδυάζονται για τη μελέτη. Δεδομένου ότι η μέθοδος χρησιμοποιεί την κατάταξη των γονιδίων σε κάθε σειρά, αντί για την πραγματική αξία της έκφρασης, μπορεί να εφαρμοστεί ευέλικτα σε πολλά διαφορετικά ζητήματα, όπως ο προσδιορισμός των γονιδίων που είναι ταξινομημένα με βάση μια κατάσταση , ενώ είναι ταξινομημένα από μια άλλη κατάσταση.

1.3.1 Μεθοδολογία αλγόριθμου RankProd.

Ο αλγόριθμος RankProd κάνει ένα είδος μετα-ανάλυσης δεδομένων συνδυάζοντας τα δεδομένα από τις διάφορες λίστες που του δίνονται με συγκεκριμένη μεθοδολογία. Για καλύτερη κατανόηση του αλγορίθμου τον σπάμε σε 5 βήματα. Έστω δύο σύνολα δεδομένων από διαφορετικές πηγές (διαφορετικά πειραματικά εργαστήρια ή στην περίπτωση μας διαφορετικούς αλγόριθμους εύρεσης υποψήφιων γονιδίων που σχετίζονται σε μία ασθένεια) T και C που προήλθαν από δύο διαφορετικές προϋποθέσεις η μία ενάντια στην άλλη και έχουμε τις n_T και n_C επαναλήψεις που αποτελούν το πρώτο σύνολο δεδομένων και τις m_T και m_C που αποτελούν το δεύτερο σύνολο δεδομένων.

(i) Για κάθε μονοδιάστατο πίνακα, υπολογίζεται ο λόγος FC των ζευγαριών σε κάθε σύνολο δεδομένων.

$$\frac{T_{n1}}{C_{n1}}, \frac{T_{n1}}{C_{n2}}, \dots, \frac{T_{nT}}{C_{nC}} \Rightarrow n_T \, x \, n_C \text{ (μεταθέσεις)}$$

$$\frac{T_{m1}}{C_{m1}}, \frac{T_{m1}}{C_{m2}}, \dots, \frac{T_{mT}}{C_{mC}} \Rightarrow m_T \, x \, m_C \text{ (μεταθέσεις)}$$

(ii) Ιεραρχεί το λόγο που έκανε σε κάθε μετάθεση. (η υψηλότερη = ιεράρχηση 1)

$$\Longrightarrow r_{gi}$$

όπου r_{gi} η ιεράρχηση του g-οστού γονιδίου με βάση την i-οστή μετάθεση, i=1..K όπου:

$$K = (n_T \, x \, n_C) + (m_T \, x \, m_C)$$

(iii) Υπολογισμός του RP με βάση τον τύπο:

$$RP_G = \prod_i r_{gi}^{1/K}$$

(iv) Ανεξάρτητα μεταθέτει την τιμή της έκφρασης ή της ιεράρχησης του γονιδίου σε κάθε πίνακα σε σχέση με το ID του γονιδίου και επαναλαμβάνει τα πιο πάνω βήματα από (i)-(iii)

$$\Longrightarrow RP_g^{(l)}$$

(v) Επαναλαμβάνεται το βήμα (iv) L φορές, σχηματίζοντας αναφορές με κατανομή:

$$RP_g^{(i)}(l = 1, \dots, L)$$

και καθορίζει το p-value και FDR που σχετίζεται με κάθε γονίδιο.

1.3.2 Συναρτήσεις RankProd.

Συνάρτηση topGene: Η συνάρτηση αυτή παρουσιάζει τα ευρισκόμενα γονίδια μαζί με τις συναφής τους στατιστικές, σε δύο πίνακες, τα υπέρ-εκφρασμένα και τα υπο-εκφρασμένα γονίδια στην κλάση 1 σε σύγκριση με την κλάση 2.(Εικόνα 1.3.1) Ο χρήστης έχει την ικανότητα να θέση τα γονίδια που επιστρέφει η συνάρτηση αυτή να είναι μικρότερη ή μεγαλύτερη τιμή από το ποσοστό ψευδώς θετικών προβλέψεων (pfp). Οι πίνακες που επιστρέφει, στο πίνακας 1 η κλάση 1 < κλάσης 2 και παρουσιάζει τα υπέρ-εκφρασμένα γονίδια ενώ ο πίνακας 2 η κλάση 1> κλάσης 2 και παρουσιάζει τα υπό-εκφρασμένα γονίδια. Ο κάθε πίνακας έχει 4 στήλες, στην πρώτη είναι τα ονόματα των γονιδίων, στην δεύτερη είναι το υπολογισμένο RP για κάθε γονίδιο, στην τρίτη είναι η υπολογιζόμενη αλλαγή της κατανομής (κλάση 1/κλάση 2) στα επίπεδα του κατά μέσου όρου έκφρασης με βάση τις δύο προϋποθέσεις , οι οποίες μετατρέπονται στο αρχικό μέγεθος χρησιμοποιώντας τα δεδομένα σε λογάριθμο με βάση το 2 και στην τελευταία το ποσοστό της ψευδός θετικής πρόβλεψης που συγκέντρωσε το γονίδιο.

Συνάρτηση plotRP: Αυτή η συνάρτηση χρησιμοποιείται για να παραστήσουμε γραφικά την εκτιμώμενη ψευδός θετική πρόβλεψη (pfp) σε σχέση με τον αριθμό των γονιδίων που μας επέστεψε η συνάρτηση RP. Ο μέγιστος αριθμός αποδεκτών «pfp» αναγνωρίζεται και τα γονίδια αυτά σημαδεύονται με κόκκινο. Αυτά που δεν είναι αποδεκτά σημαδεύονται με μαύρο χρώμα. Αναπαράγονται δύο γραφικές μία για τα υπέρ-

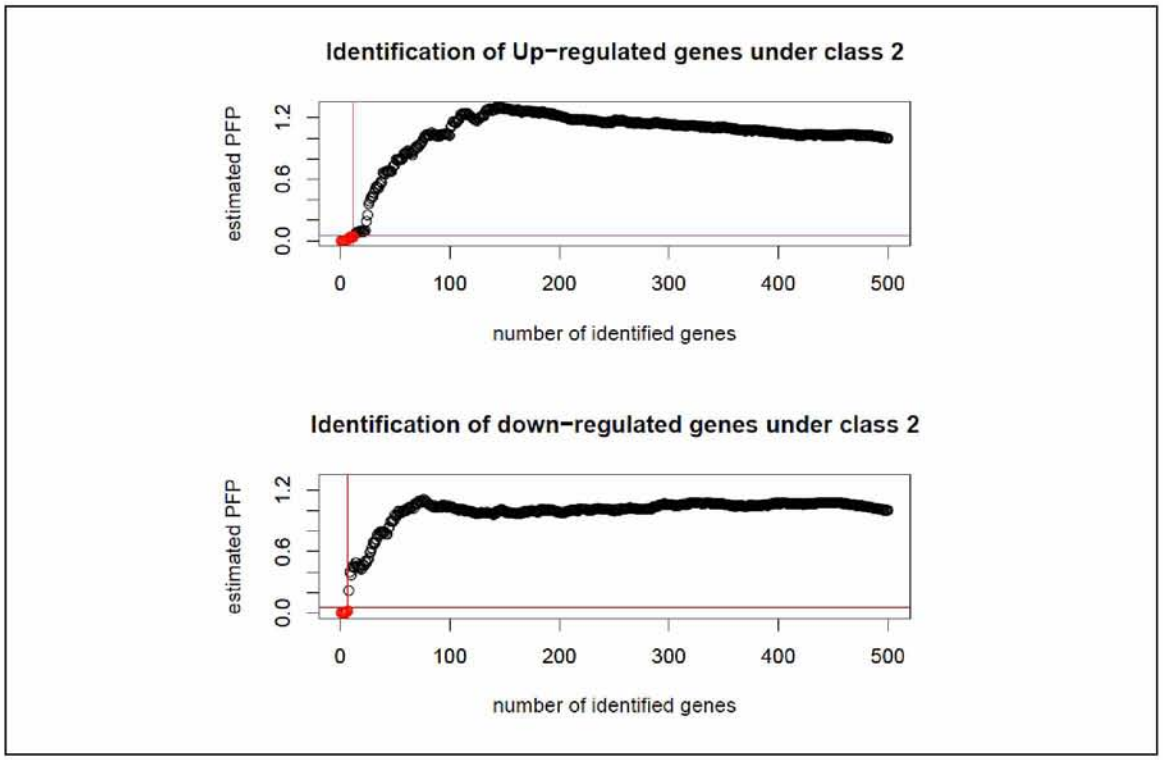
εκφραμσένα γονίδια και μία για τα υπό-εκφρασμένα γονίδια που βρίσκονται κάτω από την 2 κλάση.(Εικόνα 1.3.2)

Συνάρτηση plotGene: Αυτή η συνάρτηση χρησιμοποιείται για να παραστήσουμε γραφικά το γονίδιο που μας ενδιαφέρει με βάση τις τιμές έκφρασης του γονίδιου και τα στατιστικά αποτελέσματα που συγκέντρωσε. (Εικόνα 1.3.3.)

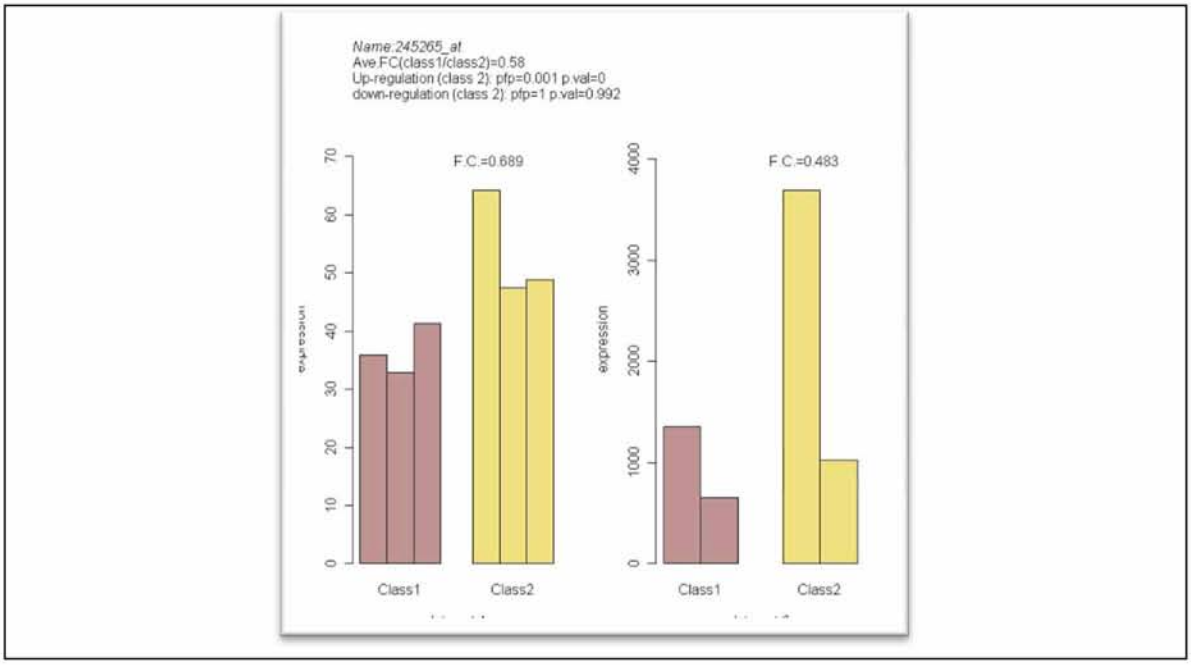
Εικόνα 1.3.1: Αποτελέσματα συνάρτησης topGene.

```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2
$Table1
gene.index RP/Rsum FC:(class1/class2) pfp
245244_at 344 1.4891 0.4327 0.0000
245336_at 436 2.4231 0.4773 0.0000
245119_at 219 3.0932 0.4783 0.0000
245176_at 276 3.3955 0.5038 0.0000
245304_at 404 3.7745 0.5011 0.0000
245196_at 296 7.9884 0.6035 0.0100
245254_at 354 9.4769 0.6469 0.0143
245262_at 362 11.0043 0.6667 0.0187
245334_at 434 14.9425 0.6994 0.0389
245141_at 241 15.2589 0.6971 0.0380
245265_at 365 15.7394 0.6888 0.0391
245112_at 212 15.7589 0.7112 0.0358
$Table2
gene.index RP/Rsum FC:(class1/class2) pfp
245362_at 462 1.0000 2.5935 0.0000
245136_at 236 2.8470 1.7180 0.0000
245277_at 377 4.9437 1.5636 0.0000
245296_at 396 5.6434 1.5505 0.0000
245276_at 376 8.5368 1.4795 0.0060
245229_at 329 9.1905 1.4577 0.0083
245075_at 175 11.7001 1.3937 0.0229
```

Εικόνα 1.3.2: Αποτελέσματα συνάρτησης plotRp.



Εικόνα 1.3.3: Αποτελέσματα συνάρτησης plotGene.



1.4 Αλγόριθμος στατιστικής ανάλυσης Metradisc.

Μεγάλη πρόκληση παρουσιάζει, ο συνδυασμός των αποτελεσμάτων από διάφορες μεγάλες σειρές δεδομένων των πολυδιάστατων βιολογικών σημάτων (όπως η σκιαγράφηση της έκφρασης των γονιδίων). Οι μεθοδολογίες αυτές είναι απαραίτητες γιατί μπορούν να συνδυάσουν αποτελεσματικά διαφορετικά είδη δεδομένων, αλλά επίσης και να ελέγξουν την έκταση της ποικιλομορφίας (ετερογένειας) σε όλες τις συνδυασμένες μελέτες. Ο αλγόριθμος Metradisc[2] παρέχει μετα-ανάλυση των συνόλων δεδομένων. Είναι μια γενίκευση της μεθόδου μετα-ανάλυσης συνδυάζοντας πληροφορίες σε όλα τα είδη δεδομένων. Για περεταίρω έλεγχο μεταξύ κάθε μελέτης που σχετίζεται με την ετερογένεια σε κάθε βιολογική παράμετρο που μας ενδιαφέρει. Ο αλγόριθμος Metradisc στηρίζεται στο μη παραμετρικό έλεγχο μετάθεσης Μόντε Κάρλο. Ανάλογα με το επίπεδο της στατιστικής σημαντικότητας ταξινομούνται οι δοκιμές των βιολογικών μεταβλητών.

Για κάθε βιολογική μεταβλητή που μας ενδιαφέρει ελέγχει το μέσο όρο ταξινόμησης και την μελέτη της ετερογένειας μεταξύ των ειδικών ιεραρχήσεων. Μετά από συνυπολογισμό των δεσμών και των διαφορών μεταβλητών που ελέγχονται σε όλες τις μελέτες, ανταλλάσσει τυχαία την ιεράρχηση των γονιδίων της κάθε μελέτης και υπολογίζει τις προσομοιωμένες μετρήσεις του μέσου όρου και της ετερογένειας. Η διαδικασία αυτή επαναλαμβάνεται για την παραγωγή μηδενικών τιμών στις μετρήσεις. Ο αλγόριθμος Metradisc είναι ένα νέο εργαλείο για το συνδυασμό πολύπλοκων συνόλων δεδομένων που προέρχονται από μαζικές δοκιμές και την εξέταση της πολυμορφίας των αποτελεσμάτων σε ολόκληρες τις συνδυασμένες μελέτες.

Παρέχει μετα-ανάλυση των ανακτώμενων ιεραρχημένων συνόλων δεδομένων που έχει αναζητήσει κάποιος ερευνητής γύρω από μία βιολογική παράμετρο, όπως τα γονίδια που πιθανός να σχετίζονται με μία ασθένεια. Είναι μια γενικευμένη μη παραμετρική μέθοδος, που χρησιμοποιείται για να συνδυάσει πληροφορίες από όλα τα ευρισκόμενα σύνολα δεδομένων που έχουμε. Ο αλγόριθμος αυτός αποτελεί γενικευμένη επέκταση της μεθόδου HEGESMA[124], μία μέθοδος μετα-ανάλυσης του γονιδιώματος που βασίζεται στην ετερογένεια. Η όλη διαδικασία βασίζεται στο μη παραμετρικό έλεγχο Μόντε Κάρλο που βασίζεται στις μεταθέσεις.

1.4.1 Υλικό και Μέθοδοι.

Δεδομένα εισόδου του αλγόριθμου Metradisc είναι οι αναλύσεις τυπικών δειγμάτων για ένα μεγάλο αριθμό βιολογικών μεταβλητών που έχουμε εξάγει από τους διάφορους αλγόριθμους αναζήτησης που διατίθενται ελεύθερα στο διαδίκτυο. Τα αποτελέσματα που εξάγει ο αλγόριθμος, παρέχουν για κάθε ελεγμένη μεταβλητή ένα στατιστικό έλεγχο και μία αντίστοιχη στατιστική σημασία ή κάποια άλλη μετρούμενη τιμή. Για παράδειγμα, στη εφαρμογή των μικροσυστοιχιών (microarrays), τα πρότυπα έκφρασης των γονιδίων για τους προσβεβλημένους σε ασθένειες ιστούς μπορεί να συγκριθούν με κανονικούς ιστούς ελέγχου[125]. Κατά την ανάλυση των αποτελεσμάτων των μικροσυστοιχιών, για κάθε γονίδιο γίνετε η εκτίμηση του ανάλογα με το πόσο έχει διαφοροποιημένη έκφραση μεταξύ του ιστού που φέρει ασθένεια και του φυσιολογικού ιστού, χρησιμοποιώντας μια γενίκευση του t-test με προσαρμογή σε πολλαπλές συγκρίσεις.[88]

Στις διάφορες δοκιμές ακολούθως εφαρμόζονται κατάλληλοι μετασχηματισμοί όπως λογαριθμικοί μετασχηματισμοί και ομαλοποίηση. Το τελικό αποτέλεσμα είναι συνήθως μία συνολική βαθμολογία σκοραρίσματος (p-value), αν και άλλες μετρήσεις μπορούν να ληφθούν υπόψη όπως η q-τιμή και άλλες τιμές που παρουσιάζονται στην έξοδο τον αποτελεσμάτων του αλγόριθμου που η κάθε μία ξεχωριστά αντιπροσωπεύει διαφορετικούς ελέγχους. Οι υπό δοκιμή βιολογικές μεταβλητές μπορούν να ταξινομηθούν σε κάθε ανάλυση με βάση την κατεύθυνση των αποτελεσμάτων (υπο-εκφρασμένα, υπερ-εκφρασμένα) και το επίπεδο της στατιστικής σημαντικότητας ή άλλες αριθμητικές μεταβλητές..

Όταν συνδυάζονται διάφορες μελέτες, είναι πιθανό ότι ο τρόπος βαθμολόγησης του p-value των γονιδίων να μην εξάγεται με ακριβώς το ίδιο βιολογικό τρόπο σε κάθε μελέτη. Ακόμα μπορεί να έχουν δοκιμαστεί με διαφορετικό αριθμό βιολογικών μεταβλητών. Ωστόσο, κάποιες μεταβλητές (γονίδια) μπορεί να είναι εμφανίζονται συχνά (επικάλυψη) από όλες τις μελέτες, ενώ άλλες μπορεί να είναι κοινές σε όλες τις υπό εξέταση μελέτες. Πριν συνδυαστούν αυτές οι μελέτες, οι μη επεξεργασμένες ιεραρχήσεις θα πρέπει να προσαρμόζονται με το μέγιστο αριθμό των ελεγχόμενων μεταβλητών (nmax) σε όλες τις συνδυασμένες μελέτες. Συνεπώς, για την i-οστή μελέτη όπου ο αριθμός των ελεγχόμενων μεταβλητών είναι (ni), η σειρά ιεράρχησης ισούται με:

$$\text{Ranks} = \text{nmax}/\text{ni}$$

1.4.2 Μέσος όρος και μετρήσεις ετερογένειας.

Ο μέσος βαθμός ιεράρχησης (R *), καθώς η μετρούμενη ανομοιογένεια (Q *) για κάθε βιολογική μεταβλητή από όλες τις μελέτες, υπολογίζεται με βάση το προσαρμοσμένο βαθμό ιεράρχησης. Η τιμή του * R, ορίζεται ως εξής:

$$R^{*}=\frac{\sum_{i=1}^{s} Ri}{s}$$

όπου Ri είναι ο βαθμός ιεράρχησης των υπό έρευνα βιολογικών μεταβλητών για τη μελέτη i (i = 1 έως s μελέτες).

Ο βαθμός Q* ορίζεται ως το άθροισμα των αποκλίσεων στο τετράγωνο για κάθε ιεράρχηση της μελέτης για τις βιολογικές μεταβλητές που μας ενδιαφέρουν από το μέσο βαθμό ιεράρχησης για την εν λόγω μεταβλητή. Είναι μία γενίκευση της Cochran Q στατιστικής και ορίζεται ως:

$$Q^{*} = \sum_{i=1}^{s} (Ri - R^{*})^{2}$$

Η σημασία της τιμής * R και Q * της κάθε βιολογικής μεταβλητής αξιολογείται εμπειρικά κατά τη κατανομή του μέσου όρου και από τις μετρήσεις ετερογένειας, υπό την μηδενική υπόθεση ότι οι τάξεις γίνονται τυχαία. Ο αλγόριθμος Metradisc αναζητεί τον εντοπισμό βιολογικών μεταβλητών που έχουν είτε πολύ υψηλή μέση ιεράρχηση ή πολύ χαμηλό μέσο όρο ιεραρχήσεων. Αυτό είναι σημαντικό κάθε φορά που υπάρχει μια κατεύθυνση στο αποτέλεσμα. Μας ενδιαφέρει ο εντοπισμός των γονιδίων που υπερεκφράζονται με μία ασθένεια, όπως καθώς και γονίδια που υποεκφράζονται με μία ασθένεια, αγνοώντας την κατεύθυνση του αποτελέσματος στην κατάταξη και το συνδυασμό των δεδομένων γιατί μπορεί να δώσει ψευδή αποτελέσματα.

Σε αντίθεση με τις περισσότερες εφαρμογές μετα-ανάλυσης, δεν εκτελείται μόνο για να εξετάσει κάποιο στατιστικό επίπεδο σημαντικότητας αλλά και το κατά πόσο οι παρατηρούμενες τιμές μεταξύ των μελετών ετερογένειας είναι πολύ υψηλές. Στατιστικώς η σημαντικά περιορισμένη ετερογένεια μπορεί να υποδεικνύει ότι τα αποτελέσματα των διαφόρων μελετών είναι ιδιαίτερα συνεπείς μεταξύ τους για τις συγκεκριμένες βιολογικές μεταβλητές, παρά τις διαφορές που υπάρχουν στις συνδυασμένες μελέτες. Όταν αυτό κρίνεται για μια συγκεκριμένη βιολογική μεταβλητή (γονίδιο) που συνέβαλε σημαντικά υψηλό ή σημαντικά χαμηλό μέσο βαθμό, η πολύ χαμηλή ετερογένεια μπορεί να ερμηνεύεται ως περαιτέρω απόδειξη για τη σημασία της.

Αντίθετα, πολύ μεγάλη ετερογένεια υποδηλώνει ότι τα αποτελέσματα είναι ασυμβίβαστα μεταξύ των διαφορετικών μελετών.

Η αριθμητική ετερογένεια μπορεί ορισμένες φορές να εξαρτάται και από τη μέση ιεράρχηση της κάθε βιολογικής μεταβλητής (γονίδιο)[126]. Έτσι, όταν η σημαντική ετερογένεια βρεθεί, πρέπει να ελεγχθεί κατά πόσον η ετερογένεια είναι επίσης σημαντική όταν εξετάζουμε μόνο τις βιολογικές μεταβλητές (γονίδια) που έχουν παρόμοια μέση κατάταξη.

1.4.2 Η χρήση του Monte Carlo ελέγχου στο αλγόριθμο MetraDisc.

Προκειμένου να εκτιμηθούν οι χαρακτηριστικές τιμές των μετρήσεων R^*και Q^* χρησιμοποιείται η μέθοδος Μόντε Κάρλο. Με την μέθοδο Μόντε Κάρλο μεταθέτονται τυχαία τα ιεραρχημένα γονίδια και υπολογίζονται οι υποθετικές μετρήσεις R^*και Q^*. Μετά επαναλαμβάνεται η όλη διαδικασία για να παραχθούν οι κατανομές υπό την μηδενική υπόθεση(μηδενικές υποθέσεις) για τις μετρήσεις. Ακολούθως κάθε μεταβλητή ελέγχεται ξανά με την μηδενική υπόθεση αντίστοιχα με τις πληροφορίες της ίδιας κλάσης. Για παράδειγμα βιολογικές μεταβλητές (γονίδια) με πληροφορίες από ένα συγκεκριμένο αριθμό μελετών ελέγχονται ξανά στην μηδενική υπόθεση ως παράγωγα (συμπληρώματος) από τις συγκεκριμένες έρευνες. Αν οι βιολογικές μεταβλητές (γονίδια) είχαν καταμετρηθεί «k» από «s» μελέτες, μετά οι μετρήσεις R^*και Q^* ελέγχονται ξανά με την μηδενική υπόθεση από όλες τις υποθετικές μεταβλητές δεδομένων, όπου οι παραλλαγές παρεμβάλλουν δεδομένα από αυτές τις ίδιες «k» μελέτες, αλλά χάνουν δεδομένα από τις άλλες s-k μελέτες.

Ο αριθμός των κλάσεων εξαρτάται από τον αριθμό των ερευνών και τις πόσες χαμένες περιοχές δημιουργούνται κατά την δημιουργία του πίνακα με τις κλάσεις. Για $3^{\eta\varsigma}$ έρευνες, μπορούμε να δημιουργήσουμε το πολύ 7 κλάσεις. Αυτό βγαίνει από τις υποθέσεις: (i) τα δεδομένα είναι διαθέσιμα για όλες τις μελέτες, (ii) τα δεδομένα είναι διαθέσιμα μόνο από την μία μελέτη, (iii) τα δεδομένα είναι διαθέσιμα από την δεύτερη μελέτη, (iv) τα δεδομένα είναι διαθέσιμα από την τρίτη μελέτη, (v) τα δεδομένα είναι διαθέσιμα από τις μελέτες ένα και δύο, (vi) τα δεδομένα είναι διαθέσιμα από τις μελέτες ένα και τρία, (vii) τα δεδομένα είναι διαθέσιμα από τις μελέτες δύο και τρία. Γενικά για S μελέτες ο μέγιστος πιθανός αριθμός κλάσεων που θα έχουμε είναι:

$$K = 2^S + a - 1$$

όπου το a αντιπροσωπεύει τον αριθμό των κενών περιοχών που δημιουργούνται στον πίνακα με τις κλάσεις μας.

Ο αριθμός των παραλλαγών που θα τρέξουν για να παράγονται οι μηδενικές υποθέσεις εξαρτάται από την απαιτούμενη ακρίβεια στις τελικές τιμές (p-value) της μετα-ανάλυσης. Γενικά με «F» διαθέσιμες προσομοιώσεις η ακρίβεια στις πληροφορίες της κάθε κλάσης δεν μπορεί να υπερβαίνει το 1/Fg όπου g είναι ο αριθμός των βιολογικών μεταβλητών (γονίδια) που έχει η κάθε κλάση.

Το σημαντικό επίπεδο για τις υψηλές ιεραρχήσεις σε μέσο όρο R^* ορίζεται ως το ποσοστό των προσομοιωμένων μετρήσεων οι οποίες υπερβαίνουν ή είναι ισοδύναμες στο παρατηρούμενο $R^*(R^*observed)$. Το στατιστικό σημαντικό επίπεδο για το περιορισμένο ποσοστό ετερογένειας Q^* είναι το ποσοστό των μετρήσεων που είναι λιγότερο ή ίσο με το παρατηρημένο $R^*(R^*observed)$. Το στατιστικό σημαντικό επίπεδο για την υψηλή ετερογένεια είναι το ποσοστό των προσομοιωμένων μετρήσεων που είναι υψηλότερες ή ίσες στο παρατηρούμενο $Q^*(Q^*observed)$.

1.4.3 Η χρήση σταθμισμένων και αστάθμιστων ερευνών στην μετα-ανάλυση.

Η προεπιλεγμένη ανάλυση του αλγόριθμου είναι για να απονέμει ίσα βάρη σε όλες τις μελέτες(αστάθμιστη ανάλυση). Εναλλακτικά, μπορούμε να δώσουμε βάρη σε κάθε έρευνα, πολύ απλά από το μέγεθος της ή με άλλες συναρτήσεις εύρεσης του βάρους που αντιστοιχεί στην κάθε μελέτη. Μετά από έρευνα και πειραματικές δοκιμές αποδείχτηκε ότι οι ετερογενετικοί έλεγχοι είναι καλύτερο να δείξουν την διαφορετικότητα των μικρών ερευνών ενάντια στις μεγάλες έρευνες. Ο μέσος ιεραρχικός έλεγχος ενδέχεται να είναι προτιμότερος για την διεξαγωγή μίας αστάθμιστης ανάλυσης ειδικά όταν οι συνδεδεμένες μελέτες χρησιμοποιούν πολύ διαφορετικά δεδομένα, από διαφορετικές εμπειρικές συνθήκες ή ακόμη και σε διαφορετικά είδη, και έτσι δεν υπάρχει λογική σύνδεση των δεδομένων, αφού τα αποδεικτικά στοιχεία μπορεί να είναι διαφορετικά από κάποιες άλλες.

Πέρα από αυτό μερικές φορές οι έρευνες που εμπλέκονται στις μετα-αναλύσεις περιέχουν πολύ διαφορετικά ποσά δεδομένων, με διακύμανση από πολύ μικρή μέχρι πολύ μεγάλη. Σε αυτή την περίπτωση η σταθμευμένη ανάλυση θα μας επιφέρει καλύτερα αποτελέσματα.

Τότε η σταθμισμένη μέση ιεράρχηση Rw^* για κάθε βιολογική μεταβλητή διαμέσου των ερευνών ορίζεται ως:

$$Rw^* = \sum_{i=1}^{s} wiRi \; / \; \sum_{i=1}^{s} wi$$

όπου το Ri η ιεράρχηση της βιολογικής μεταβλητής(γονίδιο) κάτω από την έρευνα της μελέτης i (από i μέχρι s μελέτες) και wi είναι το σχετικό βάρος για κάθε έρευνα. Η συνάρτηση στάθμισης έχει ανάγκη να καθορίζεται κάθε φορά, ανάλογα από το είδος των δεδομένων που συνδυάζονται. Όταν τα δεδομένα αναφέρονται σε στατιστικές παράγονται από το t-test έλεγχο συγκρίνοντας τους μέσους όρους των δύο ομάδων. Μία καλή συνάρτηση στάθμισης ορίζεται ως:

$$Wi = (ni1 * ni2)/(ni1+ni2)$$

όπου ni1 και ni2 είναι ο αριθμός των γονιδίων που δεν σχετίζονται με ασθένειες και των γονιδίων που σχετίζονται με μία ασθένεια στην έρευνα i αντίστοιχα. Όταν ο αριθμός των κανονικών γονιδίων και των γονιδίων που σχετίζονται με ασθένειες είναι ο ίδιος τότε απλοποιούνται τα βάρη και είναι ίδια για όλες τις έρευνες.

ΚΕΦΑΛΑΙΟ 2: ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΛΟΓΙΑΣ

2.1.Αναζήτηση υποψήφιων γονιδίων.

Η αναζήτηση των υποψήφιων γονιδίων είναι μία διαδικασία όπου πρέπει να γίνει σχολαστικά και επίμονα για κάθε ασθένεια σε κάθε αλγόριθμο αναζήτησης. Ο ορισμός των παραμέτρων πρέπει να γίνεται σωστά ούτως ώστε να έχουμε τα καλύτερα αλλά και πιο αξιόπιστα αποτελέσματα. Σε αυτό το κομμάτι οι παράμετροι που χρησιμοποιούμε για κάθε μας αναζήτηση στους εννέα διαφορετικούς αλγόριθμους αναζήτησης είναι οι λέξεις κλειδιά .

Οι λέξεις κλειδιά είναι οι χαρακτηριστικές λέξεις που αντιπροσωπεύουν την κάθε ασθένεια ξεχωριστά όπως ορίζονται από παγκόσμιο οργανισμό υγείας με βάση την διεθνή ταξινόμηση των νόσων (ICD-9)[23] . Η Disease Ontology (DO)[22] είναι μία μεγάλη βάση δεδομένων που περιέχει μέσα όλες τις ονοματολογίες , συνώνυμες λέξεις κλειδιά που συσχετίζονται με κάθε ασθένεια. Με την βοήθεια των παραπάνω βάσεων-κανονισμών κάνουμε μια αναζήτηση ευρείας κλίμακας ούτως ώστε να εντοπίσουμε όλες τις πιθανές λέξεις κλειδιά για τις έξι ασθένειες που θα ασχοληθούμε. Στον πίνακα 2.1.1 που ακολουθεί παρουσιάζονται όλες οι λέξεις κλειδιά που βρήκαμε για κάθε ασθένεια ξεχωριστά.

Μετά από έρευνα διαπιστώσαμε ότι ο καθένας από τους διαφορετικούς μας αλγόριθμους δέχεται τις διάφορες λέξεις κλειδιά με διαφορετικό τρόπο. Έτσι για να είμαστε σίγουροι ότι θα εξάγουμε τα καλύτερα αποτελέσματα αναζήτησης από κάθε αλγόριθμο δοκιμάσαμε όλες τις λέξεις κλειδιά μας ακόμη και εναλλάσσοντας τις λέξεις που απαρτίζουν την κάθε λέξει κλειδί. Επίσης μετά από αρκετές αναζητήσεις σε κάθε αλγόριθμο ξεχωριστά διαπιστώσαμε ότι για κάθε συνώνυμη λέξει κλειδί είναι δυνατό να μας παρουσιάζει διαφορετικά αποτελέσματα, είτε λιγότερα , είτε περισσότερα κάθε φορά.

Έτσι μετά από κάθε αναζήτηση σε κάθε αλγόριθμο, ήταν δυνατό να έχουμε από μία μέχρι πέντε λίστες με υποψήφια γονίδια που σχετίζονταν για κάθε ασθένεια ξεχωριστά. Η κάθε μία λίστα που είχαμε με βάση τις λέξεις κλειδιά που εισάγαμε, αντιπροσώπευε κάθε ασθένεια και πιθανός περιείχε κάποια κοινά υποψήφια γονίδια ή κάποια διαφορετικά υποψήφια γονίδια ανάλογα από των αριθμό των συνώνυμων λέξεων-κλειδιών που είχαμε για κάθε ασθένεια . Όλες οι λίστες περιείχαν τα υποψήφια γονίδια που σχετίζονταν για κάθε ασθένεια ταξινομημένα με ιεραρχική σειρά ανάλογα με την

σειρά επιτυχίας τους, με κριτήριο την συγκεντρωτική βαθμολογία (E-Value) που συγκέντρωνε το καθένα.

Ο τρόπος βαθμολόγησης των υποψήφιων γονιδίων ήταν διαφορετικός κάθε φορά ανάλογα με τον κάθε αλγόριθμο που χρησιμοποιούσαμε . Ο κάθε αλγόριθμος τύχαινε να έχει διαφορετικές βιολογικές παραμέτρους βαθμολόγησης, διαφορετικές στατιστικές αναλύσεις για την διεξαγωγή της συγκεντρωτικής βαθμολογίας (E-Value) του κάθε υποψηφίου γονίδιου αλλά και διαφορετικές πηγές δεδομένων για εισαγωγή πληροφοριών.

Πίνακας 2.1.1: Λέξεις κλειδιά που αντιπροσωπεύουν κάθε ασθένεια.

ΑΣΘΕΝΕΙΑ	ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ
Καρκίνος Μαστού (Cancer Breast)	-Cancer Breast -Breast Cancer -Breast Carcinoma -Carcinoma mastos -Tumor of breast -Breast Tumor -Benign Tumor of Breast
Διαβήτης Τύπου II (Diabetes Type II)	-Diabetes 2 -Diabetes II -Diabetes mellitus Type 2 -Diabetes mellitus Type II -Type II Diabetes -Type 2 Diabetes -Diabetes Type II -Diabetes Type 2 -Diabetes mellitus 2 -Diabetes mellitus II -Diabetes Mellitus, Non Insulin Dependent
Διαβήτης Τύπου I (Diabetes Type I)	-Diabetes 1 -Diabetes I -Diabetes mellitus Type 1 -Diabetes mellitus Type I -Type I Diabetes -Type 1 Diabetes -Diabetes Type I -Diabetes Type 1 -Diabetes mellitus 1 -Diabetes mellitus I -Diabetes Mellitus, Insulin Dependent

Υπέρταση **(Hypertension)**	-High blood pressure -HBP -Hypertension -Pressure -HT -HTN -HPN -blood pressure monitoring -diastolic hypertension -hypertensive heart disease -malignant hypertension -systolic hypertension -blood flow -heart attack -heart disease -arterial disease -coronary heart disease -systolic pressure -peripheral vascular disease -diastolic pressure
Παχυσαρκία **(Obesity)**	-Obesity -Adipose tissue -Overweight -Corpulence -Blubber
Κατά πλάκα Σκλήρυνση **(Multiple sclerosis)**	-Multiple sclerosis -abbreviated MS -disseminated sclerosis -encephalomyelitis disseminata -MS

2.2. Δημιουργία λίστας με γονίδια για κάθε μέθοδο/αλγόριθμο ξεχωριστά.

Η δημιουργία της συγκεντρωτικής λίστας με τα υποψήφια γονίδια για κάθε ασθένεια, σε κάθε αλγόριθμο ξεχωριστά ήταν μία αναγκαία διαδικασία που έπρεπε να διεκπεραιωθεί ούτως ώστε να έχουμε μόνο μία άρτια και εύκολα προσβάσιμη λίστα με σκοπό την γρήγορη επεξεργασία των δεδομένων της, η οποία θα περιέχει όλα τα υποψήφια γονίδια που βρέθηκαν να σχετίζονται με μία ασθένεια με όλες τις λέξεις κλειδιά που συνδέονται με την ασθένεια.

Για την δημιουργία της κάθε συγκεντρωτικής λίστας για κάθε ασθένεια για κάθε αλγόριθμο, μαζεύαμε συγκεκριμένες πληροφορίες από κάθε υπό-λιστα (αποτελέσματα που μας εξάγει ο κάθε αλγόριθμος με κάθε λέξει κλειδί που του εισάγαμε για μία ασθένεια) που ήταν τα υποψήφια γονίδια, η σειρά κατάταξη τους, και η συγκεντρωτική βαθμολογία (E-Value) που συγκέντρωνε το καθένα. Όλα αυτά τα δεδομένα τα αποθηκεύαμε σε ένα αρχείο .txt (Εικόνα 2.2.1) όπου στην συνέχεια θα τα επεξεργαζόμασταν για την παραγωγή της τελικής λίστας.

Η επεξεργασία των δεδομένων για την παραγωγή της συγκεντρωτικής λίστας για κάθε ασθένεια σε κάθε αλγόριθμο έπρεπε να γίνει με ένα συγκεκριμένο τρόπο και να έχει μία συγκεκριμένη μορφή για να μπορέσουμε να επεξεργαστούμε τα δεδομένα της στα επόμενα στάδια της ανάλυσης. Για αυτό το σκοπό αυτό αναλύαμε τα δεδομένα του αρχείου .txt στο πρόγραμμα Matlab με κάποιο πρόγραμμα που δημιουργήσαμε (Εικόνα 2.2.2) ώστε να μπορέσουμε να θέσουμε τους δικούς μας περιορισμούς.

Ένας από τους βασικός περιορισμός ήταν, η καινούργια συγκεντρωτική λίστα να περιέχει όλα τα υποψήφια γονίδια της κάθε υπό-λίστας ταξινομημένα με βάση την συγκεντρωτική βαθμολογία (E-Value) που είχε το καθένα σε ιεραρχικά σειρά. Σε περίπτωση που κάποιο γονίδιο εντοπιζόταν πάνω από δύο φορές στην συγκεντρωτική λίστα τότε κρατούσαμε αυτό που είχε την μεγαλύτερη γενική βαθμολογία (E-Value) και τα υπόλοιπα αφαιρούνταν από την λίστα.

Τέλος η καινούργια συγκεντρωτική λίστα αποθηκευόταν σε μορφή .txt και αποτελείτο από τρείς στήλες, την σειρά κατάταξης του κάθε υποψήφιου γονιδίου, το όνομα του υποψήφιου γονιδίου και την συγκεντρωτική του βαθμολογία (E-value). (Εικόνα 2.2.3)

Εικόνα 2.2.1: Παράδειγμα αρχείου .txt με τις πληροφορίες που εισάγαμε από κάθε υπό-λιστα για τον καρκίνο του μαστού στην GeneCards.

Λέξη Κλειδί « Α»			Λέξει Κλειδί «Β»			Λέξει Κλειδί «Γ»		
BRCA1	27.98	human	ERBB2	15.95	human	NKX2-1	10.61	human
BRCA2	24.35	human	BRCA1	15.29	human	ESR1	8.41	human
ESR1	19.15	human	CDH1	13.99	human	PGR	7.54	human
ERBB2	18.43	human	ESR1	13.57	human	VHL	7.35	human
TP53	15.1	human	BRCA2	12.94	human	CYP19A1	5.26	human
ABCG2	12.35	human	MUC1	9.69	human	MUC1	5.16	human
EGFR	12.23	human	TP53	9.1	human	VIM	5.06	human
PTEN	11.1	human	CTSD	8	human	EGF	4.89	human
CDH1	10	human	ABCG2	7.46	human	PLAU	4.73	human
CTSD	9.78	human	TFF1	6.16	human	PCNA	4.6	human
MLH1	8.36	human	EGFR	5.96	human	MMP2	4.38	human
TFF1	7.64	human	VEGFA	5.66	human	PTEN	4.35	human
TGFB1	7.33	human	PGR	5.66	human	MKI67	4.34	human
KLK3	7.27	human	GRP	5.6	human	MMP9	4.33	human
MSH2	7	human	SNCG	5.56	human	CTSD	4.32	human
CTNNB1	6.69	human	BCL2	5.48	human	TGFB1	4.25	human
PGR	6.62	human	ERBB3	5.36	human	KLK3	4.15	human
GSTM1	6.58	human	KLK10	5.28	human	FGF2	4.02	human
CYP1A1	6.51	human	CDH3	5.15	human	IGF2	3.99	human
GSTP1	6.46	human	CXCR4	5.04	human	DES	3.81	human
BRMS1	6.34	human	CCND1	4.98	human	ACTC1	3.8	human
BCAR3	6.34	human	AMPH	4.8	human	TGFA	3.78	human
CHEK2	6.23	human	PTK6	4.69	human	BMP6	3.78	human
SRC	6.21	human	BRMS1	4.67	human	TFF1	3.7	human
NRG1	6.08	human	PCNA	4.63	human	ESR2	3.56	human
RB1	6	human	CTTN	4.21	human	BIRC5	3.55	human
ERBB3	5.99	human	EGF	4.19	human	MDM2	3.5	human
TYMS	5.63	human	PTGS2	4.07	human	PLAUR	3.48	human
PIK3CA	5.61	human	CDKN1A	4.02	human	AR	3.45	human
CTAG1B	5.54	human	CTNNB1	3.98	human	IGFBP3	3.39	human

Εικόνα 2.2.2: Κώδικας υλοποιημένου ενδεικτικού προγράμματος σε Matlab για παραγωγή συγκεντρωτικής λίστας για τον καρκίνο του μαστού στην GeneCards.

```
fid=fopen('Gene_Cards_Cancer Breast_List_all.txt');
A=textscan(fid,'%s');
j=1;   N=size(A{1});
for i=1:N
    a=strfind(A{1}(i),'human');
    if ~isempty(a{1})
        fprintf('%s\n',A{1}{i-1});
        S(j).name = A{1}{i-2};
        S(j).score = A{1}{i-1};
        j=j+1;
      end
end
fclose(fid);
```

```
[m,n]=size(S);
D=[];
for i=1:n
    s= S(i).score;
    D(i,1) = sscanf(s,'%f');
    D(i,2)=i;
end
[m,n]=size(D);
for i=1:m-1
for j=1:m-1
if D(j,1) < D(j+1,1)
    o=D(j,1);
    p=D(j,2);
D(j,1)=D(j+1,1);
D(j,2)=D(j+1,2);
D(j+1,1)=o;
D(j+1,2)=p;
end
end
end
fprintf('TAXINOMHMENA\n\n');
S(100000).name='louis';
for i=1:m
    for j=1:m-i
    e=D(i,2);
    f=D(i+j,2);
    eh=S(e).name;
    fh=S(f).name;
    z=sscanf(eh,'%c');
    x=sscanf(fh,'%c');
    [mo,no]=size(z);
    [ko,lo]=size(x);
    if ((mo==ko) && (no==lo))
         if z == x
         D(i+j,2)=10000;
         end
        end
    end
end
metritis=1;
for i=1:m
    e=D(i,2);
    if e == 100000
    continue
    end
    fprintf('%i \t %s \t %s \n',metritis,S(e).name,S(e).score);
    metritis = metritis+1;
end
```

Εικόνα 2.2.3: Παράδειγμα αρχείου .txt με μία από τις τελικές συγκεντρωτικές λίστες για την ασθένεια Καρκίνου του μαστού στον αλγόριθμο GeneCards.

```
1       BRCA1       27.98
2       BRCA2       24.35
3       ESR1  19.15
4       ERBB2       18.43
5       TP53  15.1
6       CDH1  13.99
7       ABCG2       12.35
8       EGFR  12.23
9       PTEN  11.1
10      NKX2-1      10.61
11      CTSD  9.78
12      MUC1  9.69
13      MLH1  8.36
14      TFF1  7.64
15      PGR   7.54
```

2.3 Δημιουργία τελικής λίστας με γονίδια και βάρη για κάθε ασθένεια.

Μετά από αρκετές αναζήτησής και των 9 αλγόριθμων που χρησιμοποιούμε στην μελέτη μας βρισκόμαστε στην ανάγκη να δημιουργήσουμε μία τελική λίστα που να περιέχει μέσα όλα τα γονίδια από τις «τελικές λίστες αλγόριθμων» (λίστες που πήραμε μέσα από την επεξεργασία στο πρόγραμμα μας στην Matlab) με τα γονίδια που βρέθηκαν να σχετίζονται με την εκάστοτε ασθένεια που αναζητούμε μέσα από τους 9 διαφορετικούς αλγόριθμους αναζήτησης.

Η δημιουργία αυτής της λίστας επείγει για να μπορέσουμε να εισάγουμε μέσα περισσότερους περιορισμούς και βαθμονόμηση σε κάθε γονίδιο που περιέχει η λίστα. Δηλαδή, ανάλογα με τις πόσες φορές εμφανίζεται (αν περιέχεται και στις 9 «τελικές λίστες αλγόριθμων» που αντιπροσωπεύουν η κάθε μία τον κάθε αλγόριθμο ξεχωριστά), ανάλογα με την θέση που εμφανίζεται και στις 9 «τελικές λίστες αλγόριθμων» και στη ποινή που θα πρέπει να προστεθεί στα βάρη του κάθε γονιδίου στην περίπτωση που δεν βρίσκεται μέσα σε μία από τις 9 «τελικές λίστες αλγόριθμων». Επίσης ένα μείζων σημασίας θέμα στην βαθμονόμηση μας είναι η τιμή του βάρους που θα παίρνει το κάθε γονίδιο ανάλογα με την θέση ιεράρχησης που επίτυχε στην κάθε «τελική λίστα αλγορίθμου» αλλά και το πόσο μεγάλο εύρος σε γονίδια αναζητεί ο κάθε αλγόριθμος για να παράγει τα αποτελέσματα του. Για παράδειγμα τα γονίδια που παρουσιάζονται πολύ ισχυρά να συνδέονται με μία ασθένεια του αλγόριθμου Α, που εκτελεί την αναζήτηση του σε 5.000 γονίδια θα πρέπει να τους δίνονται μεγαλύτερα βάρη από τα γονίδια που παρουσιάζονται να είναι πάλι πολύ ισχυρά συνδεδεμένα με μία ασθένεια του αλγόριθμου Β που εκτελεί την αναζήτηση του σε 3.000 γονίδια.

Η όλοι αυτοί διαδικασία γίνεται για να μπορέσουμε να εισάγουμε πλέον την τελική λίστα με όλα τα γονίδια και τα βάρη τους που σχετίζονται με μία ασθένεια, σε άλλους αλγόριθμους που λειτουργούν με βάση στατιστικές αναλύσεις, και να μας παρουσιάσουν με την σειρά τους τα πιο υψηλά υποψήφια γονίδια που θα σχετίζονται με μία ασθένεια.

2.3.1 Βαθμονόμηση με βάρη του κάθε γονίδιου.

Η όλη διαδικασία βαθμονόμησης με βάρη του κάθε γονιδίου εκτελείται σε πρόγραμμα στην Matlab και παρουσιάζεται στην Εικόνα 2.3.1 . Ως δεδομένα εισόδου το πρόγραμμα μας δέχεται ένα αρχείο .txt το οποίο περιέχει μέσα και της 9 «τελικές λίστες

αλγορίθμων». Για προσδιορισμό του αλγόριθμου αναζήτησης που παράχθηκαν τα δεδομένα της κάθε λίστας προσθέτουμε δίπλα από την σειρά επιτυχίας και το όνομα του κάθε γονιδίου, το όνομα του αλγόριθμου από τον οποίο προήλθε. Για παράδειγμα στην «τελική λίστα αλγόριθμου» του αλγόριθμου Gene Prospector τα γονίδια της λίστας σημαδεύονται με «GenePr». Στο πίνακα 2.3.1 παρουσιάζεται παράδειγμα με τα δεδομένα που εισάγονται στο πρόγραμμα μας στο Matlab για την ασθένεια καρκίνου του μαστού.

Η βαθμονόμηση των γονιδίων μετά εκτελείται με τον ακόλουθα βήματα:

(i) Μαζεύουμε όλα τα γονίδια από το αρχείο μας, ξεκινώντας την ανάγνωση των γονιδίων στήλη με στήλη και κρατάμε για το καθένα δείκτες που μας προσδιορίζουν της παραμέτρους name(όνομα), score(σειρά κατάταξης) και base(σε πιο αλγόριθμο αναζήτησης ανήκει το κάθε γονίδιο). Με βάση το παράδειγμα του πίνακα 2.3.1 αρχικά κρατούνται τα ονόματα από τα γονίδια της πρώτης στήλης που ανήκουν στον αλγόριθμο PosMed μετά τα ονόματα των γονιδίων της δεύτερης στήλης που ανήκουν στον αλγόριθμο SNPs3D, ούτω καθεξής μέχρι την τελευταία στήλη.

(ii) Συγκρίνουμε όλα τα γονίδια μεταξύ τους από την αρχή μέχρι το τέλος ως προς την ονομασία τους. Αν κάποιο γονίδιο εμφανίζεται πάνω από δύο φορές τότε οι περισσότερες από μια εμφανίσεις του σημαδεύονται με μία χαρακτηριστική τιμή. Με βάση το πρόγραμμα μας που παρουσιάζεται στην Εικόνα 2.3.1 τα γονίδια που εμφανίζονται περισσότερες από μία φορές σημαδεύονται με την τιμή 10000. Ο τρόπος με τον οποίο σαρώνονται τα δεδομένα, αξίζει να σημειωθεί ότι πρώτα συγκρίνονται τα δεδομένα της πρώτης στήλης (δηλαδή του αλγόριθμου PosMed) και μετά αν κάποιο από τα γονίδια των επόμενων στηλών δεν υπάρχει στην πρώτη στήλη τότε ακολουθεί ως επόμενο μη σημαδεμένο γονίδιο. Τέλος δημιουργείται μία καινούργια λίστα με όλα τα μη σημαδεμένα γονίδια και έτσι έχουμε μία λίστα που να περιέχει όλα τα γονίδια από το αρχείο .txt μόνο μία φόρα.

Πίνακας 2.3.1: Δεδομένα που εισάγονται σε πρόγραμμα στην Matlab για βαθμονόμηση κάθε γονιδίου και δημιουργίας της τελικής λίστας με γονίδια και βάρη.

Τα δεδομένα που παρουσιάζονται πιο κάτω είναι τυπικό δείγμα με τα πρώτα 15 γονίδια από τα 820 γονίδια που εισήχθηκαν στο πρόγραμμα για επεξεργασία.

1	ST18	POSMED		1	BRCA1	SNP3ds		1	SRC	CANDID		1	PPP2R1B	FITSNP		1	BRCA1	GenePr		1	MUC16	Phenob		1	ERBB2	Suspec		1	NP_055483	Prospe		1	BRCA1	GeneCa
2	BCKDHB	POSMED		2	VDR	SNP3ds		2	GRB2	CANDID		2	AURKA	FITSNP		2	BRCA2	GenePr		2	JAK2	Phenob		2	CYP19A1	Suspec		2	NP_149081	Prospe		2	BRCA2	GeneCa
3	HRES1	POSMED		3	ERBB2	SNP3ds		3	PIK3R1	CANDID		3	PTEN	FITSNP		3	TP53	GenePr		3	ABL1	Phenob		3	ESR1	Suspec		3	ABCG2	Prospe		3	ESR1	GeneCa
4	AMPH	POSMED		4	TP53	SNP3ds		4	TP53	CANDID		4	NCOA3	FITSNP		4	GSTM1	GenePr		4	ESR1	Phenob		4	AKR1B1	Suspec		4	NCOA3	Prospe		4	ERBB2	GeneCa
5	ZYX	POSMED		5	BRCA2	SNP3ds		5	PTPN11	CANDID		5	AKAP13	FITSNP		5	GSTT1	GenePr		5	BRCA1	Phenob		5	MSH6	Suspec		5	BIN2	Prospe		5	TP53	GeneCa
6	BRCA2	POSMED		6	PKM2	SNP3ds		6	FYN	CANDID		6	CCND1	FITSNP		6	ESR1	GenePr		6	PIK3R1	Phenob		6	FZR1	Suspec		6	BCAR3	Prospe		6	CDH1	GeneCa
7	TRERF1	POSMED		7	EGFR	SNP3ds		7	TGFBR1	CANDID		7	MDM2	FITSNP		7	CYP1A1	GenePr		7	PML	Phenob		7	KLK3	Suspec		7	NP_789783	Prospe		7	ABCG2	GeneCa
8	BRCA1	POSMED		8	VEGFA	SNP3ds		8	SHC1	CANDID		8	ATM	FITSNP		8	CYP1B1	GenePr		8	PPP2CA	Phenob		8	ATM	Suspec		8	PPIL4	Prospe		8	EGFR	GeneCa
9	BCAR3	POSMED		9	CYP17A1	SNP3ds		9	EGFR	CANDID		9	IGFBP7	FITSNP		9	FGFR2	GenePr		9	PRKCA	Phenob		9	CDH1	Suspec		9	CHRDL2	Prospe		9	PTEN	GeneCa
10	BCAS4	POSMED		10	CYP1A2	SNP3ds		10	PTPN6	CANDID		10	MYC	FITSNP		10	ATM	GenePr		10	CHEK1	Phenob		10	RAD51	Suspec		10	SNCG	Prospe		10	NKX2-1	GeneCa
11	BCAS1	POSMED		11	DAAM1	SNP3ds		11	PLCG1	CANDID		11	RB1	FITSNP		11	MTHFR	GenePr		11	PMS2	Phenob		11	XRCC3	Suspec		11	NP_071353	Prospe		11	CTSD	GeneCa
12	GREB1L	POSMED		12	CYP3A5	SNP3ds		12	UBC	CANDID		12	MSH6	FITSNP		12	CHEK2	GenePr		12	RAP1A	Phenob		12	PTEN	Suspec		12	BCAR1	Prospe		12	MUC1	GeneCa
13	BRMS1L	POSMED		13	FMNL2	SNP3ds		13	RAF1	CANDID		13	GHRHR	FITSNP		13	XRCC1	GenePr		13	FRK	Phenob		13	TSG101	Suspec		13	Q8NBT9	Prospe		13	MLH1	GeneCa
14	BRMS1	POSMED		14	FADS2	SNP3ds		14	AKT1	CANDID		14	RAD50	FITSNP		14	COMT	GenePr		14	MYC	Phenob		14	PHB	Suspec		14	SEPT1	Prospe		14	TFF1	GeneCa
15	FAM84B	POSMED		15	FMNL1	SNP3ds		15	STAT3	CANDID		15	NBN	FITSNP		15	CYP17A1	GenePr		15	AKT1	Phenob		15	RAD54L	Suspec		15	PTK6	Prospe		15	PGR	GeneCa

(iii) Μαζί με την ονομασία του κάθε γονιδίου υπολογίζονται και κρατούνται 9 τιμές που αντιπροσωπεύουν τα βάρη που χαρακτηρίζουν κάθε γονίδιο.

$$\sum_{i=1}^{n} Gi \;\; Weight1\, Weight2\, Weight3 \ldots Weight9$$

όπου G είναι το όνομα του γονιδίου, και n το συνολικό μέγεθος σε γονίδια της λίστας που δημιουργήθηκε πιο πάνω. Τα 9 χαρακτηριστικά βάρη που ακολουθούν αντιπροσωπεύουν τις 9 στήλες με τα γονίδια που περιείχε το αρχείο μας που εισάγαμε, όπου η κάθε στήλη περιέχει τα γονίδια του κάθε αλγόριθμου ξεχωριστά. Στο πίνακα 2.3.2 αναλύονται ξεχωριστά σε πιο αλγόριθμο αναζήτησης αντιστοιχεί το κάθε βάρος αλλά και πώς υπολογίζεται η τιμή του κάθε βάρους.

Πίνακας 2.3.2: Επεξήγηση τρόπου εύρεσης τιμής κάθε βάρους ξεχωριστά αλλά και σε πιο αλγόριθμο αναζήτησης αντιστοιχεί.

A/A	Πηγάζει από τα δεδομένα αλγόριθμου	Υπολογισμός		Πλήθος γονιδίων αναζήτησης (A)
		Κανονικά	Με Ποινή	
W1	POSTMED	*W1= A1-J1*	*W1= (A1-JT1)/2*	A1= 39928
W2	SNPs3D	*W2= A2-J2*	*W2= (A2-JT2)/2*	A2= 39928
W3	CANDID	*W3= A3-J3*	*W3= (A3-JT3)/2*	A3= 39928
W4	FITSNPs	*W4= A4-J4*	*W4= (A4-JT4)2*	A4= 39928
W5	GENEPROSPECTOR	*W5= A5-J5*	*W5= (A5-JT5)/2*	A5= 39928
W6	PHENOPRED	*W6= A6-J6*	*W6= (A6-JT6)/2*	A6= 39928
W7	SUSPECTOR	*W7= A7-J7*	*W7= (A7-JT7)/2*	A7= 39928
W8	PROSPECTR	*W8= A8-J8*	*W8= (A8-JT8)/2*	A8= 3739
W9	GENECARDS	*W9= A9-J9*	*W9= (A9-JT9)/2*	A9= 39928

Με βάση τον πίνακα 2.3.2 η τιμή του κάθε βάρους υπολογίζετε η κανονικά ή με ποινή. Για να εκτελεστή το βάρος κανονικά πρέπει το γονίδιο που αναζητούμε να αναγνωριστεί μέσα στις στήλες του αρχείου μας που εισάγουμε. Η θέση που αναγνωρίζεται αλλά και η στήλη που βρίσκεται αντιπροσωπεύει την παράμετρο J. Δηλαδή αν ένα γονίδιο αναγνωριστεί στην 5 στήλη στην θέση 10 θέση, τότε το J5 θα πάρει την τιμή 10. Όσα γονίδια δεν αναγνωρίζονται από κάποιες στήλες του αρχείου μας, σημαίνει ότι το εν λόγο γονίδιο που αναζητούμε στο αρχείο μας δεν βρέθηκε να συσχετίζει με την εν λόγο ασθένεια που μελετούμε σε κάποιο από τους αλγόριθμους

αναζήτησης μας. Επομένως θα πρέπει να βάλουμε ποινή (faoul) στο γονίδιο, στην θέση της συγκεκριμένης στήλης του αρχείου μας που δεν βρέθηκε.

Στον πίνακα 2.3.3 που ακολουθεί παρουσιάζεται τμήμα από την τελική λίστα με γονίδια και βάρη που έχουμε βρει για την ασθένεια καρκίνο του μαστού. Ο πίνακας αυτός συμβαδίζει με τα δεδομένα του πίνακα 2.3.1 , όπου εύκολα κανείς μπορεί να τους συνδυάσει και να καταλάβει καλύτερα πως λειτουργεί η όλη διαδικασία με τα βάρη.

Πίνακας 2.3.3: Τμήμα με αποτελέσματα από την τελική λίστα με βάρη και γονίδια της ασθένειας καρκίνου του μαστού.

Α/Α	Γονίδια	Βάρος 1	Βάρος 2	Βάρος 3	Βάρος4	Βάρος5	Βάρος 6	Βάρος 7	Βάρος8	Βάρος 9
1	ST18	39927	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
2	BCKDHB	39926	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
3	HRES1	39925	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
4	AMPH	39924	1.99E+04	39545	19755	19939	1.99E+04	1847	2.00E+04	39874
5	ZYX	39923	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
6	BRCA2	39922	39891	39742	39923	39926	1.99E+04	3722	39901	39926
7	TRERF1	39921	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
8	BRCA1	39920	39904	39892	39927	39927	39923	3723	39903	39927
9	BCAR3	39919	1.99E+04	19714	19755	19939	1.99E+04	3733	2.00E+04	39903
10	BCAS4	39918	1.99E+04	19714	39585	19939	1.99E+04	3695	2.00E+04	19554
11	BCAS1	39917	1.99E+04	19714	39594	19939	1.99E+04	3709	2.00E+04	39817
12	GREB1L	39916	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
13	BRMS1L	39915	1.99E+04	19714	19755	19939	1.99E+04	3704	2.00E+04	39592
14	BRMS1	39914	1.99E+04	19714	39832	19939	1.99E+04	3715	2.00E+04	39904
15	FAM84B	39913	1.99E+04	19714	39758	19939	1.99E+04	1847	2.00E+04	19554

Εικόνα 2.3.1: Κώδικας του προγράμματος όπου δημιουργεί την καινούργια τελική λίστα με βάρη για την κάθε ασθένεια σε γλώσσα προγραμματισμού Matlab.

```
fid=fopen('name off file.txt');
A=textscan(fid,'%s');
j=1;
N=size(A{1});
faulPostmed=0;
for i=1:N
    a=strfind(A{1}(i),'POSMED');
    if ~isempty(a{1})
        S(j).score=A{1}{i-2};
        S(j).name = A{1}{i-1};
        S(j).base= A{1}{i};
        j=j+1;
        faulPostmed=faulPostmed+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
B=textscan(fid,'%s');
N=size(B{1});
faulFitsnp=0;
for i=1:N
    b=strfind(B{1}(i),'FITSNP');
    if ~isempty(b{1})
        S(j).score=B{1}{i-2};
        S(j).name = B{1}{i-1};
        S(j).base= B{1}{i};
        j=j+1;
        faulFitsnp=faulFitsnp+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
C=textscan(fid,'%s');
N=size(C{1});
faulCandid=0;
for i=1:N
    c=strfind(C{1}(i),'CANDID');
    if ~isempty(c{1})
        S(j).score=C{1}{i-2};
        S(j).name = C{1}{i-1};
        S(j).base= C{1}{i};
        j=j+1;
        faulCandid=faulCandid+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
D=textscan(fid,'%s');
N=size(D{1});
faulSNP=0;
for i=1:N
    d=strfind(D{1}(i),'SNP3ds');
    if ~isempty(d{1})
```

```
        S(j).score=D{1}{i-2};
        S(j).name = D{1}{i-1};
        S(j).base= D{1}{i};
        j=j+1;
        faulSNP=faulSNP+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
E=textscan(fid,'%s');
N=size(E{1});
faulGenePr=0;
for i=1:N
    e=strfind(E{1}(i),'GenePr');
    if ~isempty(e{1})
        S(j).score=E{1}{i-2};
        S(j).name = E{1}{i-1};
        S(j).base= E{1}{i};
        j=j+1;
        faulGenePr=faulGenePr+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
F=textscan(fid,'%s');
N=size(F{1});
faulPhenob=0;
for i=1:N
    f=strfind(F{1}(i),'Phenob');
    if ~isempty(f{1})
        S(j).score=F{1}{i-2};
        S(j).name = F{1}{i-1};
        S(j).base= F{1}{i};
        j=j+1;
        faulPhenob=faulPhenob+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
F=textscan(fid,'%s');
N=size(F{1});
faulProspe=0;
for i=1:N
    f=strfind(F{1}(i),'Prospe');
    if ~isempty(f{1})
        S(j).score=F{1}{i-2};
        S(j).name = F{1}{i-1};
        S(j).base= F{1}{i};
        j=j+1;
        faulProspe=faulProspe+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
F=textscan(fid,'%s');
```

```
N=size(F{1});
faulSuspec=0;
for i=1:N
    f=strfind(F{1}(i),'Suspec');
    if ~isempty(f{1})
        S(j).score=F{1}{i-2};
        S(j).name = F{1}{i-1};
        S(j).base= F{1}{i};
        j=j+1;
        faulSuspec=faulSuspec+1;
    end
end
fclose(fid);

fid=fopen('Rank_MS_all.txt');
F=textscan(fid,'%s');
N=size(F{1});
faulGeneCa=0;
for i=1:N
    f=strfind(F{1}(i),'GeneCa');
    if ~isempty(f{1})
        S(j).score=F{1}{i-2};
        S(j).name = F{1}{i-1};
        S(j).base= F{1}{i};
        j=j+1;
        faulGeneCa=faulGeneCa+1;
    end
end
fclose(fid);

 [m,n]=size(S);
  for i=1:n
      fprintf('%s\t%s\t%s\n',S(i).score,S(i).name,S(i).base);
  end

  D=[];
     for i=1:n
       s= S(i).score;
       D(i,1) = sscanf(s,'%f');
       D(i,2)=i;
       D(i,3)=i;
     end

fprintf('DHMIOURGIA LISTAS ME OLA TA GONIDIA APO TIS LISTES MAS\n\n');
[m,n]=size(D);
S(10000).name='louis';
    for i=1:m
        for j=1:m-i
        e=D(i,3);
        f=D(i+j,3);
        eh=S(e).name;
        fh=S(f).name;
        z=sscanf(eh,'%c');
        x=sscanf(fh,'%c');
        [mo,no]=size(z);
        [ko,lo]=size(x);
        if ((mo==ko) && (no==lo))
            if z == x
            D(i+j,3)=10000;
```

```
            end
        end
        end
    end
L=[]; %DHMIOYRGIA DEIKTH NEAS LISTAS ME TA GONIDIA APO OLES TIS LISTES
count=1;
   for i=1:m
       e=D(i,3);
        if e == 10000
        continue
        end
       fprintf('%s\n',S(e).name);
       Li(count).name= S(e).name;
       L(count,1) =count ;
       count=count+1;
   end
%DHMIOYRGIA STOIXEIVN YPOLISTON GIA KATHE ALGORITHMO
%SKORARISMA KAI GEMISMA PINAKA L ME DEDOMENA APO OLES TIS LISTES
 [m,n]=size(L);
 [k,l]=size(D);
   for i=1:m
     for j=1:k
    e=L(i,1);
     f=D(j,2);
     eh=Li(e).name;
     fh=S(f).name;
     z=sscanf(eh,'%c');
     x=sscanf(fh,'%c');
     o=S(f).base;
         if o == 'POSMED'
             [mo,no]=size(z);
             [ko,lo]=size(x);
             if ((mo==ko) && (no==lo))
             if z==x
                 L(i,2) = 39928-(faulPostmed+1-j);
             end
             end
         end
         if o == 'FITSNP'
             [mo,no]=size(z);
             [ko,lo]=size(x);
             if ((mo==ko) && (no==lo))
             if z==x
                 L(i,3) = 39928-(faulFitsnp+1-(j-faulPostmed));
             end
             end
         end
         if o == 'CANDID'
             [mo,no]=size(z);
             [ko,lo]=size(x);
             if ((mo==ko) && (no==lo))
             if z==x
                 L(i,5)      =      39928-(faulCandid+1-(j-faulPostmed-
faulFitsnp));
             end
             end
         end
         if o == 'SNP3ds'
             [mo,no]=size(z);
```

```
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,6)   =   39928-(faulSNP+1-(j-faulPostmed-faulFitsnp-
faulCandid));
            end
            end
        end
        if o == 'GenePr'
            [mo,no]=size(z);
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,7) = 39928-(faulGenePr+1-(j-faulPostmed-faulFitsnp-
faulCandid-faulSNP));
            end
            end
        end
        if o == 'Phenob'
            [mo,no]=size(z);
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,8) = 39928-(faulPhenob+1-(j-faulPostmed-faulFitsnp-
faulCandid-faulSNP-faulGenePr));
            end
            end
        end
        if o == 'Prospe'
            [mo,no]=size(z);
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,9)  =  3739-(faulProspe+1-(j-faulPostmed-faulFitsnp-
faulCandid-faulSNP-faulGenePr-faulPhenob));
            end
            end
        end
        if o == 'Suspec'
            [mo,no]=size(z);
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,10)      =      39928-(faulSuspec+1-(j-faulPostmed-
faulFitsnp-faulCandid-faulSNP-faulGenePr-faulPhenob-faulProspe));
            end
            end
        end
        if o == 'GeneCa'
            [mo,no]=size(z);
            [ko,lo]=size(x);
            if ((mo==ko) && (no==lo))
            if z==x
                L(i,11)      =      39928-(faulGeneCa+1-(j-faulPostmed-
faulFitsnp-faulCandid-faulSNP-faulGenePr-faulPhenob-faulProspe-
faulSuspec));
            end
            end
        end
```

```
     end
  end

   [m,n]=size(L);
   for i=1:m
    if L(i,2)==0
      L(i,2)=(39928-faulPostmed)/2;
    end
    if L(i,3)==0
      L(i,3)=(39928-faulFitsnp)/2;
    end
    if L(i,5)==0
      L(i,5)=(39928-faulCandid)/2;
    end
    if L(i,6)==0
      L(i,6)=(39928-faulSNP)/2;
    end
    if L(i,7)==0
      L(i,7)=(39928-faulGenePr)/2;
    end
    if L(i,8)==0
      L(i,8)=(39928-faulPhenob)/2;
    end
    if L(i,9)==0
      L(i,9)=(3739-faulProspe)/2;
    end
    if L(i,10)==0
      L(i,10)=(39928-faulSuspec)/2;
    end
    if L(i,11)==0
      L(i,11)=(39928-faulGeneCa)/2;
    end
    e=L(i,1);

fprintf('%i\t%s\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\n',L(i,1),Li(e).name
,L(i,2),L(i,3),L(i,5),L(i,6),L(i,7),L(i,8),L(i,9),L(i,10),L(i,11));

L(i,4)=L(i,2)+L(i,3)+L(i,5)+L(i,6)+L(i,7)+L(i,8)+L(i,9)+L(i,10)+L(i,11);
  end

    for i=1:m-1
      for j=1:m-1
           if L(j,4) < L(j+1,4)
        o=L(j,1);
        p=L(j,4);
    L(j,1)=L(j+1,1);
    L(j,4)=L(j+1,4);
    L(j+1,1)=o;
    L(j+1,4)=p;
           end
      end
    end
 fprintf('SKORARISMENA     ME     VASH     TO     E-VALUE     TOUS\n\n'
,L(i,1),Li(e).name,L(i,4));

 for i=1:m
    e=L(i,1);
    fprintf('%i\t%s\t%i \n',i,Li(e).name,L(i,4));
 end
```

2.4. Δημιουργία λίστας με βάρη για συσχετίσεις γονιδίων μεταξύ δύο ασθενειών.

Μέχρι τώρα οι τελικές λίστες μας με τα υποψήφια γονίδια που φτιάχναμε για να τα εισάγουμε στα προγράμματα συσχέτισης και στατιστικής ανάλυσης γονιδίων, αποτελούσαν δεδομένα για υποψήφια γονίδια που είχαμε πάρει από κάποιο αλγόριθμο αναζήτησης ή συγκεντρωτικά δεδομένα με τα υποψήφια γονίδια που βρήκαμε για μία ασθένεια με βάρη ιεράρχησης από όλους τους αλγόριθμους αναζήτησης. Σε αυτό το μέρος θα περιγράψουμε το τελευταίο στάδιο επεξεργασίας και συνδυασμού δεδομένων από δύο τελικές λίστες με βάρη και γονίδια για δύο ασθένειες που πιθανός να συνδέονται. Η καινούργια λίστα μας περιέχει τα κοινά γονίδια που παρουσίασαν στις λίστες τους οι δύο ασθένειες που θα συσχετίσουμε μαζί με τα βάρη τους (τα βάρη και από τις δύο ασθένειες) και ένα καινούργιο βάρος που θα παίρνει τιμές με βάση την παράμετρο Χ.

Για πρώτο στάδιο υπολογίζουμε την παράμετρο Χ για κάθε υποψήφιο γονίδιο από την τελική λίστα με βάρη και γονίδια για κάθε ασθένειας. Η παράμετρος Χ ισούται με:

$$X= (\sum_{i=1}^{n}(\beta i) - \kappa)/n$$

όπου i=1..n, n το πλήθος των αλγόριθμων αναζήτησης που είχαμε, β το βάρος ιεράρχησης που είχε το κάθε γονίδιο στο κάθε αλγόριθμο, και « κ » η θέση που έχει το υποψήφιο γονίδιο στην λίστα που επεξεργαζόμαστε. Στην συνέχεια υπολογίζουμε την παράμετρο Χ για όλα τα υποψήφια γονίδια και στις δύο ασθένειες ξεχωριστά.

Συνδυάζοντας της δύο τελικές λίστες με βάρη και γονίδια φτιάχνουμε μία λίστα με τα γονίδια που είναι κοινά και στις δυο λίστες και μαζί με αυτά κρατάμε στην καινούργια λίστα την τιμή της παράμετρου Χ σε δύο καινούργιες στήλες.

Γενικά η καινούργια λίστα με τα δεδομένα από τα κοινά υποψήφια γονίδια και από τις δύο ασθένειες περιέχει, στην πρώτη στήλη τα γονίδια, στην δεύτερη τη παράμετρος Χ με βάση τα βάρη της μίας ασθένειας, στην τρίτη στήλη τη παράμετρος Χ με βάση τα βάρη της άλλης ασθένειας, ακολουθούν εννέα στήλες με τα βάρη που είχε η μία ασθένεια με βάση την σειρά ιεράρχησης του κάθε γονιδίου με βάση τους εννέα αλγόριθμους αναζήτησης που ανατρέξαμε, και ακολούθως εννέα στήλες με τα βάρη της άλλης ασθένειας. (Εικόνα 2.4.1) Η όλη διαδικασία υλοποιείται εύκολα και σχετικά γρήγορα σε πρόγραμμα που έχουμε φτιάξει στην γλώσσα προγραμματισμού Matlab (Εικόνα 2.4.2).

Εικόνα 2.4.1: Παράδειγμα αρχείου .txt με μία από τις τελικές συγκεντρωτικές λίστες με όλα τα υποψήφια γονίδια που σχετίζονται σε δύο ασθένειες με τα βάρη τους.

row.names	Diabetes_1	Hypertension	POSMED_2	FITSNP_2	CANDID_2	SNP3ds_2	GenePr_2	Phenob_2	Prospe_2	Suspec_2	GeneCards_2	POSMED_1	FITSNP_1	CANDID_1	SNP3ds_1	GenePr_1	Phenob_1	Prospe_1	Suspec_1	GeneCards_1
BMPR2	2.90E+04	2.23E+04	39926	1.99E+04	1.85E+03	1.99E+04	39828	19952	39787	39902	39926	39856	19956	1.86E+03	1.99E+04	1.99E+04	19946	39734	19898	1.98E+04
MFN2	2.01E+04	1.99E+04	19564	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39925	39713	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	1.98E+04
NPPA	3.12E+04	1.99E+04	39924	1.99E+04	1.85E+03	39847	39889	19952	39303	39847	39924	39803	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	1.98E+04
NPPB	2.45E+04	2.22E+04	39917	1.99E+04	1.85E+03	39860	1.99E+04	19952	19614	19906	39923	39699	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	39819	1.98E+04
EDN1	3.14E+04	2.23E+04	39922	39911	3720	1.99E+04	3.99E+04	19952	39333	39866	39922	39871	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.98E+04
EDNRA	2.70E+04	2.22E+04	39864	3.99E+04	3.74E+03	1.99E+04	1.99E+04	19952	19614	39898	39921	39499	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.98E+04
AKT1	2.67E+04	2.90E+04	39410	1.99E+04	1.85E+03	3.98E+04	1.99E+04	19952	39913	19906	39918	39686	19956	1.86E+03	3.99E+04	1.99E+04	39894	39918	19898	3.98E+04
PIK3R1	2.45E+04	2.45E+04	39211	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39925	19906	39917	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	39924	39923	3.98E+04
EDNRB	2.45E+04	2.23E+04	39825	1.99E+04	1.85E+03	3.99E+04	1.99E+04	19952	19614	19906	39916	39726	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.98E+04
POMC	2.23E+04	2.22E+04	39472	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39915	39661	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.97E+04
APOB	2.45E+04	2.23E+04	39846	1.99E+04	1.85E+03	1.99E+04	3.99E+04	19952	19614	19906	39914	39876	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.99E+04
APOA1	2.23E+04	2.23E+04	39909	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39912	39878	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.99E+04
HMGCR	2.23E+04	2.23E+04	39485	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39911	39801	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.99E+04
NPPC	2.92E+04	1.99E+04	39735	3.99E+04	3.71E+03	3.99E+04	1.99E+04	19952	19614	39812	39909	39386	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	1.98E+04
CASP9	2.23E+04	2.23E+04	19564	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39744	19906	39907	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	39678	19898	3.99E+04
CASP8	2.45E+04	2.45E+04	39424	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39866	19906	39906	39530	19956	1.86E+03	1.99E+04	1.99E+04	19946	39841	19898	3.99E+04
CASP3	2.45E+04	2.23E+04	39480	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39892	19906	39905	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	39890	19898	3.99E+04
SP1	2.45E+04	2.22E+04	39361	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39383	19906	39904	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	39534	19898	3.96E+04
TH	2.45E+04	2.00E+04	39665	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39587	19906	39902	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	39628	19898	1.98E+04
NOS3	2.70E+04	2.45E+04	39901	3.99E+04	3.71E+03	1.99E+04	3.99E+04	19952	19614	19906	39901	39755	19956	1.86E+03	1.99E+04	3.99E+04	19946	19614	19898	3.97E+04
NOS2P1	2.01E+04	2.00E+04	19564	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39900	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.97E+04
NOS1	2.46E+04	2.45E+04	39862	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	39760	19906	39899	39813	19956	1.86E+03	1.99E+04	1.99E+04	19946	39512	19898	3.97E+04
ADORA1	2.22E+04	1.99E+04	39227	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39898	39299	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	1.98E+04
APAF1	2.01E+04	2.01E+04	19564	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39897	19564	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	3.99E+04
CRH	2.23E+04	1.99E+04	39866	1.99E+04	1.85E+03	1.99E+04	1.99E+04	19952	19614	19906	39896	39607	19956	1.86E+03	1.99E+04	1.99E+04	19946	19614	19898	1.98E+04
TNF	2.47E+04	2.68E+04	39722	3.99E+04	3.70E+03	1.99E+04	1.99E+04	19952	19614	19906	39895	39870	19956	1.86E+03	1.99E+04	3.99E+04	39922	19614	19898	3.99E+04

Εικόνα 2.4.2: Πρόγραμμα σε Matlab για δημιουργία καινούργιου πίνακα που περιέχει τα κοινά γονίδια και βάρη για τις δύο ασθένειες που εξετάζουμε μαζί με τις παράμετρους Χ της κάθε ασθένειας ξεχωριστά.

```
fid=fopen('όνομα αρχείου.txt');
A=textscan(fid,'%s');
j=1;
N=size(A{1});
faulPostmed=0;
for i=1:N
    a=strfind(A{1}(i),'AAAAA');
    if ~isempty(a{1})
        S(j).score1=A{1}{i-1};
        S(j).score2=A{1}{i-2};
        S(j).score3=A{1}{i-3};
        S(j).score4=A{1}{i-4};
        S(j).score5=A{1}{i-5};
        S(j).score6=A{1}{i-6};
        S(j).score7=A{1}{i-7};
        S(j).score8=A{1}{i-8};
        S(j).score9=A{1}{i-9};
        S(j).score10 = A{1}{i-10};
        S(j).name = A{1}{i-11};
        S(j).base= A{1}{i};

        %fprintf('%i\t%s\n',j,S(j).name);
        j=j+1;
    end
end
fclose(fid);

fid=fopen('όνομα αρχείου.txt');
A=textscan(fid,'%s');
K=1;
N=size(A{1});
faulPostmed=0;
for i=1:N
    a=strfind(A{1}(i),'BBBBB');
    if ~isempty(a{1})
        F(K).score1=A{1}{i-1};
        F(K).score2=A{1}{i-2};
        F(K).score3=A{1}{i-3};
        F(K).score4=A{1}{i-4};
        F(K).score5=A{1}{i-5};
        F(K).score6=A{1}{i-6};
        F(K).score7=A{1}{i-7};
        F(K).score8=A{1}{i-8};
        F(K).score9=A{1}{i-9};
        F(K).score10 = A{1}{i-10};
        F(K).name= A{1}{i-11};
        F(K).base= A{1}{i};
        K=K+1;
    end
end
fclose(fid);


  [m,n]=size(S);
```

```
  [mi,ni]=size(F);
  count=1;

   for i=1:n
       for j=1:ni
         eh=S(i).name;
         fh=F(j).name;
         z=sscanf(eh,'%c');
         x=sscanf(fh,'%c');
          [mo,no]=size(z);
          [ko,lo]=size(x);
          if ((mo==ko) && (no==lo))
              if z == x

fprintf('%i\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\
t%s\t%s\t%s\t%s\t%s\t%s\n',count,S(i).name,S(i).score1,F(j).score1,S(i).
score2,S(i).score3,S(i).score4,S(i).score5,S(i).score6,S(i).score7,S(i).
score8,S(i).score9,S(i).score10,F(j).score2,F(j).score3,F(j).score4,F(j)
.score5,F(j).score6,F(j).score7,F(j).score8,F(j).score9,F(j).score10);
              count=count+1;
              end
          end
       end
   end
```

2.5 Επεξεργασία δεδομένων και λειτουργία του πακέτου Rank-Prod για μία ασθένεια.

Το R-Project είναι μία γλώσσα προγραμματισμού που λειτουργεί σε ένα ειδικό περιβάλλον και παρέχει στατιστική υπολογιστική και γραφικά. Το R- Project παρέχει μια ευρεία ποικιλία από κατανομές (γραμμική και μη -γραμμική μοντελοποίηση , κλασικές κατανομές ελέγχου, σειρές ανάλυσης, ταξινόμηση , ομαδοποίηση , ...) και γραφικές τεχνικές.

Στην μελέτη μας θα χρησιμοποιήσουμε ένα πακέτο της γλώσσας R-Project που διατίθεται ελεύθερα στο κοινό μέσω διαδικτύου, το πακέτο «Rank-Prod». Το Rank-Prod είναι μία μη-παραμετρική μέθοδος για τον εντοπισμό διαφορικά εκφρασμένων (υπέρ-εκφρασμένα ή υπό-εκφρασμένα) γονιδίων με βάση το ποσοστό των εκτιμώμενων ψευδώς θετικών προβλέψεων (PfP) . Η μέθοδος μπορεί να συνδυάσει σύνολα δεδομένων από διαφορετικές αφετηρίες (μετα-ανάλυση) για να αυξηθεί η ισχύς του προσδιορισμού.

Για την ομαλή φόρτωση δεδομένων στο πακέτο Rank-Prod το αρχείο μας πρέπει να είναι σε μορφή Excel (Εικόνα 2.3.3). Το αρχείο μας στην πρώτη στήλη με όνομα «row.names» περιέχει τα ονόματα των γονιδίων που θα ιεραρχήσουμε. Στις επόμενες στήλες ακολουθούν τα χαρακτηριστικά βάρη που έχουν δημιουργηθεί με την βοήθεια του σχετικού προγράμματος στο Matlab (Εικόνα 2.4.2) για κάθε αλγόριθμο.

Μαζί με τα δεδομένα που εισάγουμε, όπως είναι σχεδιασμένος ο αλγόριθμος χρειάζεται να του ορίσουμε ακόμα τρείς παραμέτρους. Οι παράμετροι αυτοί είναι οι cl, origin και μία λίστα με τα ονόματα των γονιδίων. Η παράμετρος cl είναι η κλάση και την ορίζουμε σε κλάση 0. Ό λόγος που ορίζουμε μόνο μία κλάση είναι γιατί στην προκείμενη περίπτωση εξετάζουμε υποψήφια γονίδια που σχετίζονται με μία μόνο ασθένεια. Η παράμετρος origin ορίζεται για να γνωρίζουμε αν τα δεδομένα από κάθε λίστα έχουν κοινή προέλευση, δηλαδή από κοινό αλγόριθμο αναζήτησης, η γενικότερα από κοινή πηγή. Στην περίπτωση μας έχουμε 1-10 origins που αντιπροσωπεύουν τους διαφορετικούς αλγόριθμους που χρησιμοποιήσαμε.(Πίνακας 2.5.1). Τέλος φορτώνουμε την λίστα με τα ονόματα των γονιδίων που θα ιεραρχήσουμε στο Excel αρχείο μας. Παράδειγμα ενδεικτικού κώδικα που δημιουργήσαμε για να αναλύσουμε τα γονίδια που βρήκαμε να σχετίζονται με τον Καρκίνο του μαστού ακολουθεί στην Εικόνα 2.5.1 .

Εικόνα 2.5.1: Παράδειγμα αρχείου Excel με μερικά από τα γονίδια και τα αντίστοιχα βάρη τους όπως πρέπει να το εισάγουμε στο πακέτο στατιστικής ανάλυσης «Rank-Prod» του προγράμματος R-Project.

row.names	POSMED	FITSNP	CANDID	SNP3ds	GenePr	Phenob	Prospe	Suspec	GeneCards
ST18	39927	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
BCKDHB	39926	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
HRES1	39925	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
AMPH	39924	1.99E+04	39545	19755	19939	1.99E+04	1847	2.00E+04	39874
ZYX	39923	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
BRCA2	39922	39891	39742	39923	39926	1.99E+04	3722	39901	39926
TRERF1	39921	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
BRCA1	39920	39904	39892	39927	39927	39923	3723	39903	39927
BCAR3	39919	1.99E+04	19714	19755	19939	1.99E+04	3733	2.00E+04	39903
BCAS4	39918	1.99E+04	19714	39585	19939	1.99E+04	3695	2.00E+04	19554
BCAS1	39917	1.99E+04	19714	39594	19939	1.99E+04	3709	2.00E+04	39817
GREB1L	39916	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
BRMS1L	39915	1.99E+04	19714	19755	19939	1.99E+04	3704	2.00E+04	39592
BRMS1	39914	1.99E+04	19714	39832	19939	1.99E+04	3715	2.00E+04	39904
FAM84B	39913	1.99E+04	19714	39758	19939	1.99E+04	1847	2.00E+04	19554
BCAS3	39912	1.99E+04	19714	39511	19939	1.99E+04	3710	2.00E+04	39798
PBOV1	39911	1.99E+04	19714	19755	19939	1.99E+04	3694	2.00E+04	19554
BCAS2	39910	1.99E+04	19714	19755	19939	1.99E+04	3714	2.00E+04	39315
BCAR1	39909	1.99E+04	39775	19755	19939	1.99E+04	3727	2.00E+04	39802
ERGIC3	39908	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
SNCG	39907	1.99E+04	19714	39813	19939	1.99E+04	3729	2.00E+04	39893
TFF1	39906	1.99E+04	19714	39664	19939	1.99E+04	3719	2.00E+04	39914
JPH3	39905	1.99E+04	19714	19755	19939	1.99E+04	1847	2.00E+04	19554
ERBB2	39904	39902	39817	39925	39909	39862	1847	39927	39924

Πίνακας 2.5.1: Ορισμός των cl και origin με βάση την σειρά των παραμέτρων που έχει κάθε γονίδιο στην μελέτη μας.

Κλάση (cl)
0 0 0 0 0 0 0 0 0 0
Πηγή (origin)
1 2 3 4 5 6 7 8 9 10

Εικόνα 2.5.1: Στοιχειώδης Κώδικας για την ανάλυση δεδομένων για μία ασθένεια. Ο συγκεκριμένος κώδικας αναφέρεται στην ασθένεια Καρκίνου του μαστού.

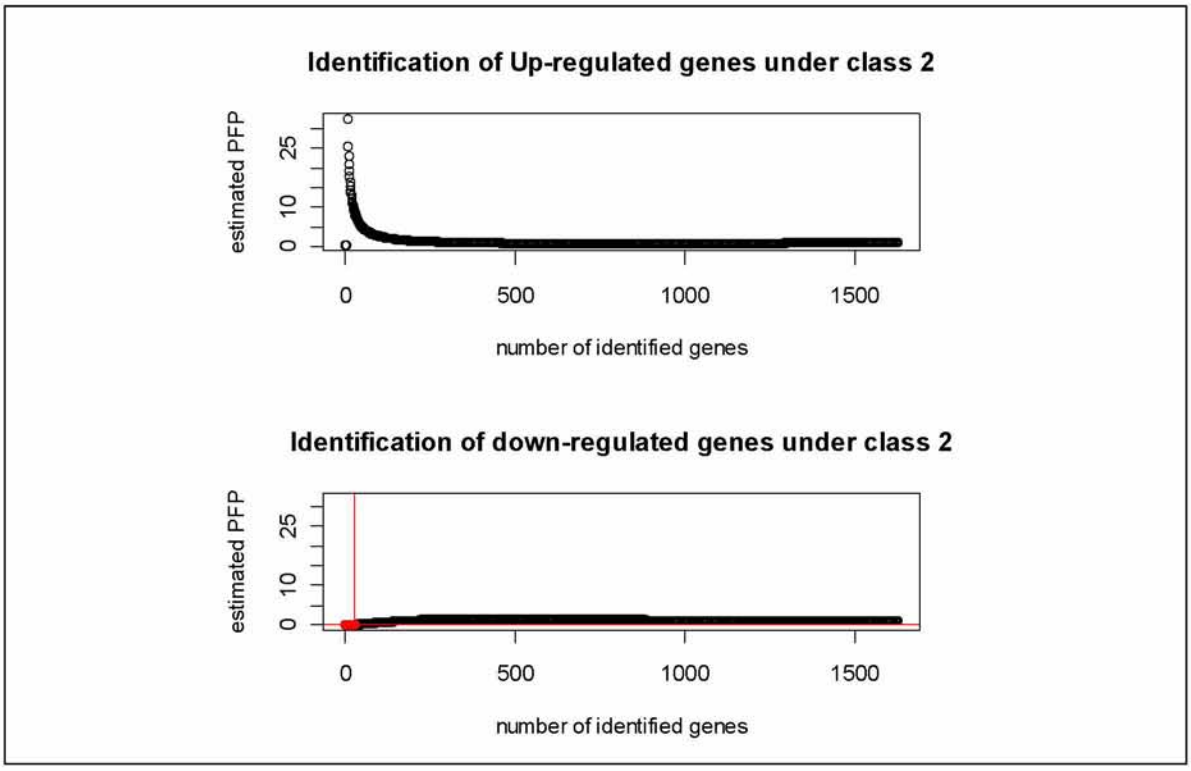
```
data1 = read.csv("c:\\Users\\Louis\\Documents\\R-Results\\
Dedomena_Breast_Cancer.csv", header = TRUE)

library(RankProd)
colnames(data1)
n1 <- 1
n2 <- 1
n3 <- 1
n4 <- 1
n5 <- 1
n6 <- 1
n7 <- 1
n8 <- 1
n9 <- 1
n10 <-1
cl <- rep(c(0,0,0,0,0,0,0,0,0,0), c(n1, n2, n3, n4, n5, n6, n7, n8, n9,
n10))
cl

origin <- rep(c(1,2,3,4,5,6,7,8,9,10), c(n1, n2, n3, n4, n5, n6, n7, n8,
n9, n10))
origin

data1.origin=origin
data1.cl=cl
```

```
data1.cl
data1.origin

A=c("ST18",
"BCKDHB",
"HRES1",
"AMPH",
"ZYX",
   '
   '
   '
"TF",
"BMP4",
"PPIG",
"ADA")
data1.gnames=A

data1.sub <- data1[, which((data1.origin == 1) | (data1.origin == 2) |
(data1.origin == 3) |(data1.origin == 4) | (data1.origin == 5) |
(data1.origin == 6) |(data1.origin == 7) | (data1.origin == 8) |
(data1.origin == 9) | (data1.origin == 10))]
data1.cl.sub <- data1.cl[which((data1.origin == 1) | (data1.origin == 2)
| (data1.origin == 3) |(data1.origin == 4) | (data1.origin == 5) |
(data1.origin == 6) |(data1.origin == 7) | (data1.origin == 8) |
(data1.origin == 9) | (data1.origin == 10))]
data1.origin.sub   <-   data1.origin[which((data1.origin   ==   1)   |
(data1.origin == 2) | (data1.origin == 3) |(data1.origin == 4) |
(data1.origin == 5) | (data1.origin == 6) |(data1.origin == 7) |
(data1.origin == 8) | (data1.origin == 9) | (data1.origin == 10))]
RP.out <- RP(data1.sub, data1.cl.sub, num.perm = 100, logged = TRUE,
na.rm = FALSE, plot = FALSE, rand = 123)
plotRP(RP.out, cutoff = 0.05)

RP.out <- RP(data1.sub,data1.cl.sub,gene.names=data1.gnames,rand=123)
RP.out <- RPadvance(data1.sub, data1.cl.sub, data1.origin.sub, num.perm
= 100, logged = TRUE, na.rm = FALSE, gene.names = data1.gnames, plot =
FALSE, rand = 123)
plotRP(RP.out, cutoff = 0.05)
```

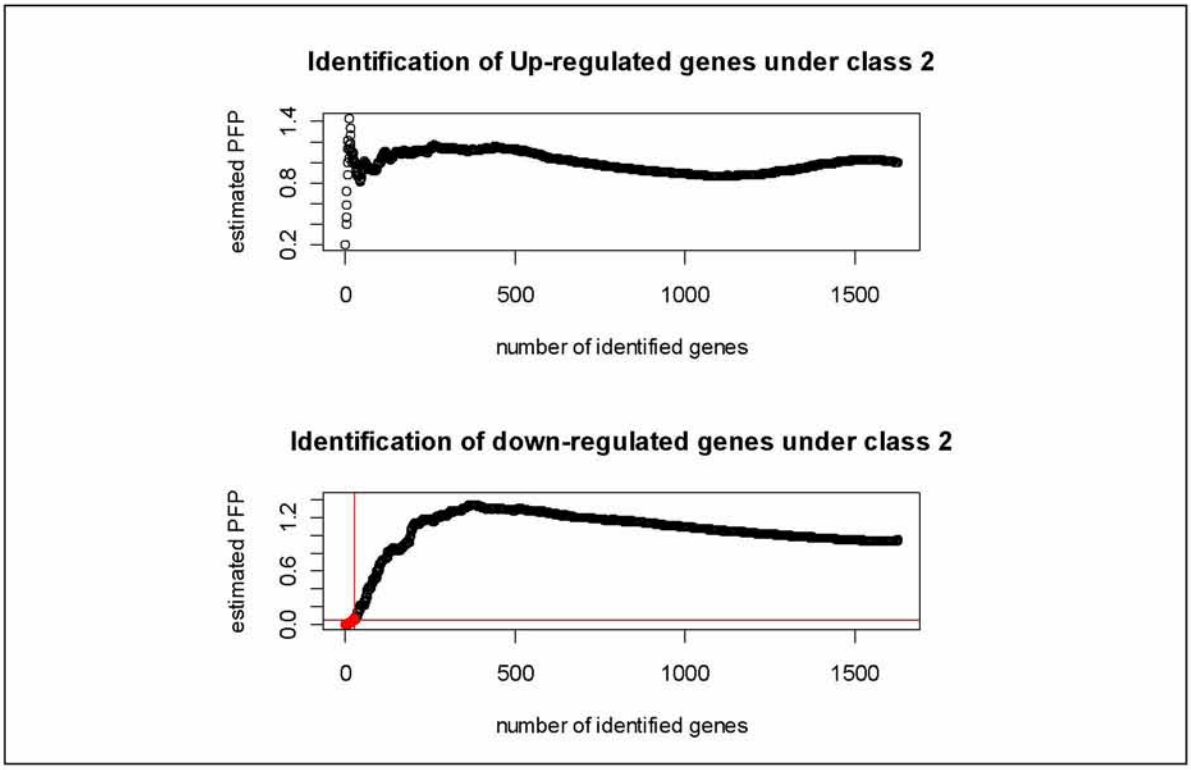
```
topGene(RP.out, cutoff = 0.05, logged = TRUE, logbase = 2, gene.names =
data1.gnames)

RP.adv.out <- RPadvance(data1, data1.cl, data1.origin, num.perm = 100,
logged = TRUE, gene.names = data1.gnames, rand = 123)
```

Με βάση τις έτοιμες συναρτήσεις που περιέχει το πακέτο του Rank-Prod όπως «RP», «plotRP», «RPadvance», «topGene» παράγουμε τα αποτελέσματα των διάφορων στατιστικών αναλύσεων. Οι συναρτήσεις που προαναφέρθηκαν περιγράφονται αναλυτικά στο κεφάλαιο 1.3. Στις εικόνες που ακολουθούν 2.5.2 και 2.5.3 παρουσιάζονται μέσω γραφικών παραστάσεων με κόκκινο χρώμα οι περιοχές στις οποίες είχαμε τα υψηλά σχετιζόμενα με την ασθένεια γονίδια με βάση την τιμή του pfp (ψευδός θετικών προβλέψεων που ορίσαμε) για τον Καρκίνο του μαστού. Επίσης τα αποτελέσματα με τα πιο υψηλά γονίδια που σχετίζονται με τον καρκίνο του μαστού παρουσιάζονται μαζί με το e-value που συγκέντρωσε το καθένα με την εκτέλεση της συνάρτησης «topGene» (Εικόνα 2.5.4). Είναι αξιοσημείωτο να σημειωθεί ότι σε όλες τις συναρτήσεις που χρησιμοποιήσαμε από το πακέτο «Rank-Prod» είχαμε θέση μεγάλο δείκτη ακρίβειας με ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05.

Εικόνα 2.5.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Καρκίνο μαστού.



Εικόνα 2.5.3: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Καρκίνο μαστού όπου RP.out=RPadvance.



Εικόνα 2.5.4: Αποτελέσματα συνάρτησης topGene() για Καρκίνο μαστού και Παχυσαρκία(δεξία) με cutoff = 0.05.

```
Table2: Genes called significant under class1 > class2

$Table1
NULL

$Table2
        gene.index  RP/Rsum FC:(class1/class2)    pfp
BRCA1            8  10.7282                 NA 0.0000
BRCA2            6  29.4108                 NA 0.0000
ERBB2           24  31.5644                 NA 0.0000
ESR1            25  35.4800                 NA 0.0000
TP53           387  35.5745                 NA 0.0000
ATM            277 108.6115                 NA 0.0017
SRC            251 135.0436                 NA 0.0129
CDH1            70 136.3081                 NA 0.0112
CYP19A1         39 140.2569                 NA 0.0111
VEGFA           31 148.8821                 NA 0.0150
VDR             86 151.7693                 NA 0.0145
EGFR            29 153.4884                 NA 0.0133
PTEN           564 153.6701                 NA 0.0123
MYC             79 156.3512                 NA 0.0121
RAD51           64 156.5159                 NA 0.0120
NCOA3          202 164.6249                 NA 0.0175
XRCC3          604 171.3363                 NA 0.0271
PIK3R1          58 172.3127                 NA 0.0267
AR             231 173.5059                 NA 0.0263
CHEK2          334 179.1722                 NA 0.0280
BARD1           45 181.4831                 NA 0.0281
CCND1          108 195.4255                 NA 0.0459
MDM2           138 202.6300                 NA 0.0565
SNCG            21 203.1116                 NA 0.0554
CDKN1A         137 204.2743                 NA 0.0552
SHBG            40 204.6497                 NA 0.0542
STAT3          154 204.9258                 NA 0.0526
CYP1A1         358 205.1102                 NA 0.0511
ABCG2          591 205.9502                 NA 0.0497
```

2.6 Επεξεργασία δεδομένων και λειτουργία του πακέτου Rank-Prod για δύο ασθένειες.

Για να μπορέσουμε να λειτουργήσουμε ομαλά χωρίς προβλήματα το πακέτο στατιστικών αναλύσεων «Rank Prod» του R-Project τα δεδομένα μας θα πρέπει να είναι αποθηκευμένα σε ένα Excel αρχείο με την μορφή όπως παρουσιάζονται στην Εικόνα 2.4.1.

Οι βασική παράμετροι που θα πρέπει να ορίσουμε στο Excel αρχείο μας είναι οι παράμετροι cl και origin και μία λίστα με τα ονόματα των γονιδίων. Η παράμετρος cl είναι η κλάση και την ορίζουμε σε κλάση 0 και κλάση 1. Αυτές οι δύο κλάσεις είναι για να γνωρίζουμε τα δεδομένα της κάθε στήλης από πια ασθένεια προέρχονται. Η παράμετρος origin ορίζεται για να γνωρίζουμε αν τα δεδομένα από κάθε λίστα έχουν κοινή προέλευση, δηλαδή από κοινό αλγόριθμο αναζήτησης, γενικότερα από κοινή πηγή. Στην περίπτωση μας έχουμε από το 1-10 origins εκ τον οποίων 9 είναι οι λίστες με τα δεδομένα που πάρθηκαν από κάθε ένα από τους εννέα διαφορετικούς αλγόριθμους που χρησιμοποιήσαμε, και το άλλο ένα που μας απομένει είναι οι λίστες που παράγονται με βάση την παράμετρο X (Πίνακας 2.6.1). Η λίστα με τα ονόματα που θα εισάγονται στον αλγόριθμο θα πρέπει να έχει την ακριβή ονοματολογία , αλλά και σειρά ιεράρχησης με τα γονίδια όπως προβάλλονται στο αρχείο Excel.

Πίνακας 2.6.1: Ορισμός των cl και origin με βάση την σειρά των παραμέτρων που έχει κάθε γονίδιο στην μελέτη μας.

Κλάση (cl)
0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
Πηγή(origin)
1 1 2 3 4 5 6 7 8 9 10 2 3 4 5 6 7 8 9 10

Αφού ορίσουμε τις παραμέτρους μας, μετά με την βοήθεια των έτοιμων συναρτήσεων που παρέχει το πακέτο «Rank Prod» που περιγράφονται αναλυτικά στο κεφάλαιο 1.3 παίρνουμε τα αποτελέσματα των διάφορων στατιστικών αναλύσεων. Ο στοιχειώδης κώδικας που αναλύει την δημιουργία κλάσεων , origins και την λίστα με τα ονόματα των γονιδίων αλλά και τον τρόπο χρήσης των διάφορων έτοιμων συναρτήσεων που χρησιμοποιούνται στο πακέτο αυτό του R-Project παρουσιάζεται στην Εικόνα 2.6.1 .

Επίσης στις εικόνες 2.6.2 και 2.6.3 παρουσιάζονται τα ενδεικτικά αποτελέσματα που είχαμε λάβει από την στατιστική ανάλυση του Διαβήτη τύπου Ι μαζί με την Παχυσαρκία. Τα αποτελέσματα με τα πιο υψηλά διαφορικά εκφρασμένα γονίδια που σχετίζονται και με τις δύο αυτές ασθένειες παρουσιάζονται και σε μορφή γραφικής παράστασης αλλά και σε πίνακα όπου αναφέρονται η ακριβής ονοματολογία των γονιδίων μαζί με το e-Value τους. Είναι αξιοσημείωτο να σημειωθεί ότι σε όλες τις συναρτήσεις που χρησιμοποιήσαμε στο R-Project είχαμε θέση μεγάλο δείκτη ακρίβειας με ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05.

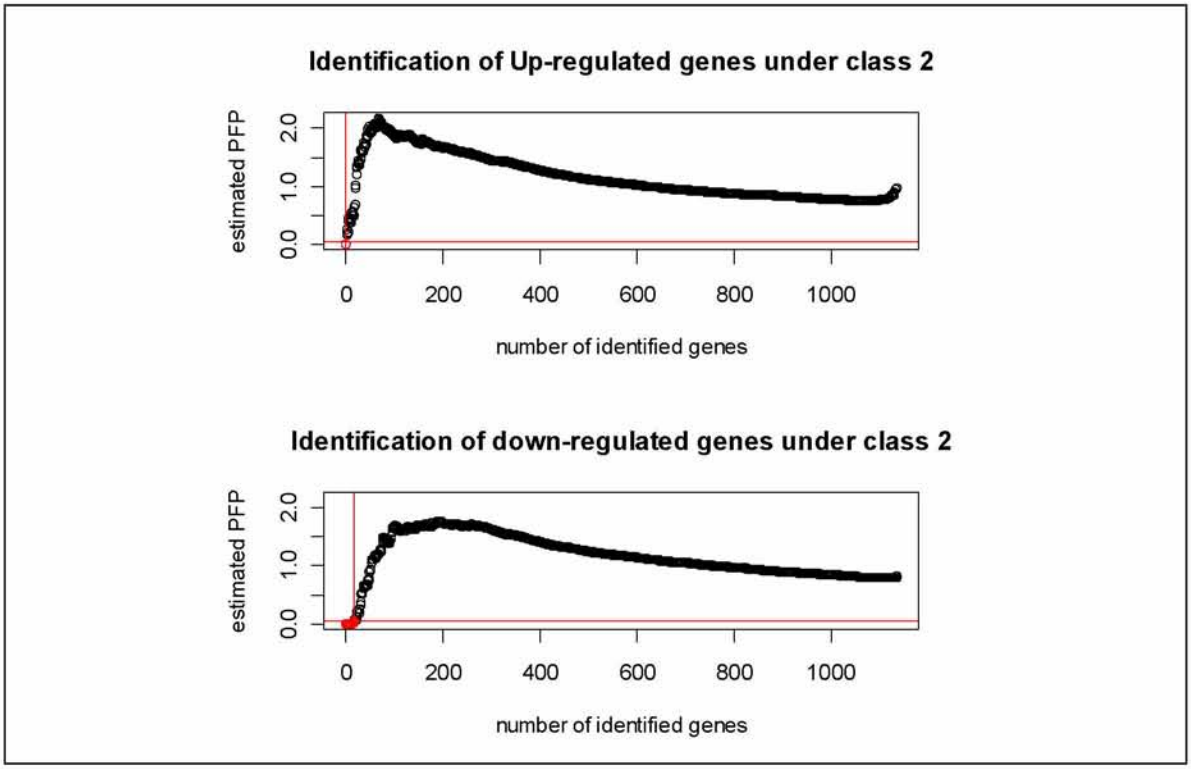
Εικόνα 2.6.1: Στοιχειώδης Κώδικας για την ανάλυση δεδομένων από δύο ασθένειες. Ο συγκεκριμένος κώδικας αναφέρεται στην ανάλυση Διαβήτη τύπου Ι μαζί με Παχυσαρκία.

```
r = read.csv("c:\\Users\\Louis\\Documents\\R-Results\\Apotelesmata_DiabetesI_Obesity.csv",
header = TRUE)
library(RankProd)
colnames(data1)
n1 <- 1
n2 <- 1
n3 <- 1
n4 <- 1
n5 <- 1
n6 <- 1
n7 <- 1
n8 <- 1
n9 <- 1
n10 <- 1
n11 <- 1
n12 <- 1
n13 <- 1
n14 <- 1
n15 <- 1
n16 <- 1
n17 <- 1
n18 <- 1
n19 <- 1
n20 <- 1

cl <- rep(c(0,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0), c(n1, n2, n3, n4, n5, n6, n7, n8, n9,n10, n11,
```

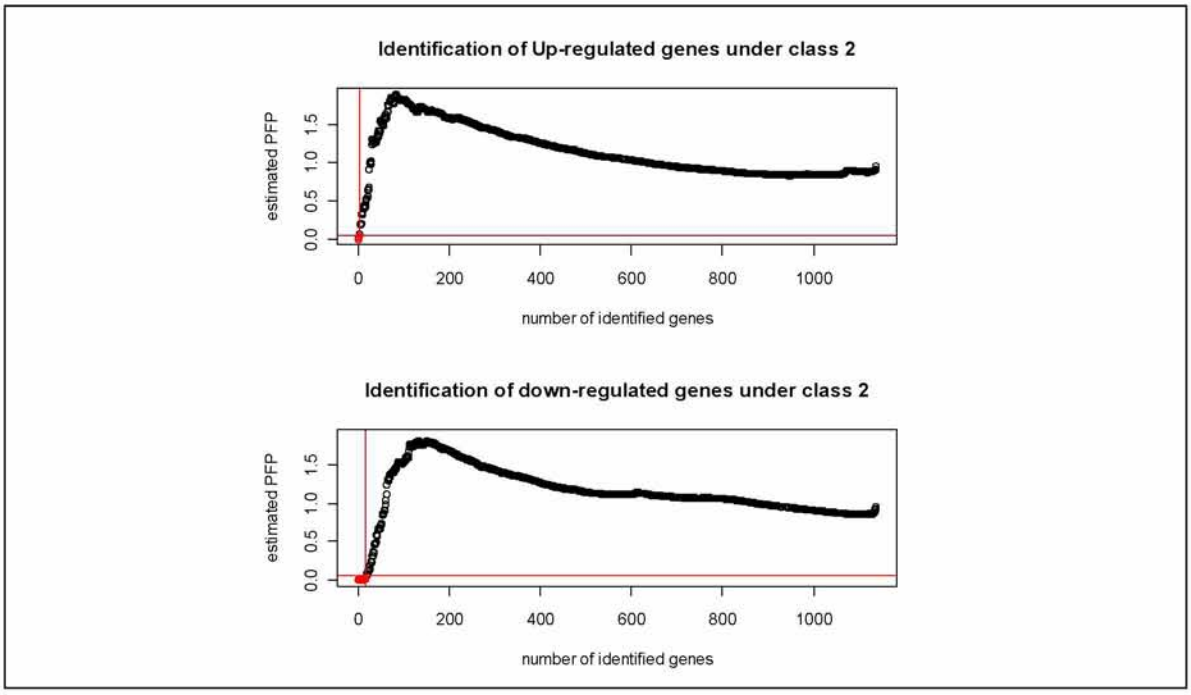
```
n12, n13, n14, n15, n16, n17, n18, n19, n20))
cl

origin <- rep(c(1,1,2,3,4,5,6,7,8,9,10,2,3,4,5,6,7,8,9,10), c(n1, n2, n3, n4, n5, n6, n7, n8, n9,n10,
n11, n12, n13, n14, n15, n16, n17, n18, n19, n20))
origin

data1.origin=origin
data1.cl=cl

data1.cl
data1.origin

A=c("GCG",
"IGF1",
"IGFBP1",
"IGF2",
"IGF1R",
 '
 '
 '
"ALPL",
"RAMP3")
data1.gnames=A

a=r[2:21]

data1 <- as.data.frame(a, row.names = data1.gnames, responseName = "Freq", stringsAsFactors
= TRUE)

data1.sub <- data1[, which((data1.origin == 1) | (data1.origin == 2) | (data1.origin == 3)
|(data1.origin == 4) | (data1.origin == 5) | (data1.origin == 6) |(data1.origin == 7) | (data1.origin ==
8) | (data1.origin == 9) | (data1.origin == 10))]
data1.cl.sub <- data1.cl[which((data1.origin == 1) | (data1.origin == 2) | (data1.origin == 3)
|(data1.origin == 4) | (data1.origin == 5) | (data1.origin == 6) |(data1.origin == 7) | (data1.origin ==
8) | (data1.origin == 9)| (data1.origin == 10))]
data1.origin.sub <- data1.origin[which((data1.origin == 1) | (data1.origin == 2) | (data1.origin == 3)
|(data1.origin == 4) | (data1.origin == 5) | (data1.origin == 6) |(data1.origin == 7) | (data1.origin ==
```

```
8) | (data1.origin == 9) | (data1.origin == 10))]

RP.out <- RP(data1.sub, data1.cl.sub, num.perm = 100, logged = TRUE, na.rm = FALSE, plot =
FALSE, rand = 123)
plotRP(RP.out, cutoff = 0.05)

RP.out <- RP(data1.sub,data1.cl.sub,gene.names=data1.gnames,rand=123)
RP.out <- RPadvance(data1.sub, data1.cl.sub, data1.origin.sub, num.perm = 100, logged =
TRUE, na.rm = FALSE, gene.names = data1.gnames, plot = FALSE, rand = 123)
plotRP(RP.out, cutoff = 0.05)

topGene(RP.out, cutoff = 0.05, logged = TRUE, logbase = 2, gene.names = data1.gnames)
```

Εικόνα 2.6.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Παχυσαρκία.



Εικόνα 2.6.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Παχυσαρκία όπου RP.out=RPadvance .



Εικόνα 2.6.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη Ι ενάντια με Σκλήρηνση κατά πλάκα με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη Ι ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Σκλήρυνσης κατά πλάκα.

```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2
$Table1
        gene.index  RP/Rsum FC:(class1/class2)   pfp P.value
PTPN22         157  45.1497                  0 0.000   0e+00
NEUROD1        136  96.6674                  0 0.015   0e+00
VDR             67 103.1377                  0 0.030   1e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
MC4R          336  34.4202                Inf 0.0000   0e+00
LIPE          373  44.3919                Inf 0.0000   0e+00
UCP3          415  46.2578                Inf 0.0000   0e+00
ADRB2         748  47.4013                Inf 0.0000   0e+00
UCP2          413  59.8925                Inf 0.0000   0e+00
PCSK1        1020  63.3182                Inf 0.0000   0e+00
AGRP          174  73.5345                Inf 0.0000   0e+00
NR3C1         423  74.5650                Inf 0.0000   0e+00
LEP            75  78.5986                Inf 0.0000   0e+00
LEPR          819  82.3356                Inf 0.0000   0e+00
POMC           89  86.6300                Inf 0.0009   0e+00
SORBS1        927  89.4212                Inf 0.0025   0e+00
PPARG         160 120.6530                Inf 0.0192   2e-04
PPARD         208 123.0689                Inf 0.0200   2e-04
GNB3          969 126.1284                Inf 0.0200   3e-04
FTO           958 129.4565                Inf 0.0200   3e-04
SREBF1        392 135.4908                Inf 0.0300   4e-04
LPL           330 137.2287                Inf 0.0306   5e-04
```

2.7 Επεξεργασία δεδομένων για εισαγωγή στο METRADISC_XL.

Με βάση τον οδηγό χρήσης του αλγόριθμου Metradisc τα δεδομένα μας πρέπει να βρίσκονται σε συγκεκριμένη μορφή, ούτως ώστε ο αλγόριθμος αυτός να μπορέσει να παρέχει με την σειρά του στατιστική ανάλυση στα δεδομένα που του εισάγουμε. Αρχικά ένα μικρό πρόβλημα που αντιμετωπίσαμε ήταν ότι στον αλγόριθμο Metradisc παρέχεται στατιστική ανάλυση σε δεδομένα που προέρχονται από επτά μελέτες.

Όπως γνωρίζεται στην παρούσα εργασία χρησιμοποιούμε εννέα διαφορετικούς αλγόριθμους για να πάρουμε τα υποψήφια γονίδια που σχετίζονται με την ασθένεια που αναζητούμε, άρα κανονικά τα δεδομένα μας προέρχονται από εννέα διαφορετικές μελέτες. Για να επιλύσουμε το παρόν πρόβλημα, συγχωνεύσαμε τα δεδομένα 2 αλγόριθμων μαζί με τα δεδομένα δύο άλλων αλγόριθμων που ο τρόπος λειτουργίας και εύρεσης των υποψηφίων γονιδίων είχε κοινά χαρακτηριστικά. Δηλαδή τα δεδομένα τα αλγόριθμου SNPs3D και του FitSNPs συγχωνεύθηκαν σε μία στήλη και του GeneProspector μαζί με τον Prospectr συγχωνεύθηκαν σε άλλη στήλη.

Γενικά ο αλγόριθμος δέχεται ως δεδομένα εισόδου δύο αρχεία, το αρχείο «data.txt» όπου μέσα είναι ένας πίνακας που περιλαμβάνει την ταυτότητα του κάθε γονιδίου (στον αλγόριθμο αυτό δεν εισάγονται τα ονόματα των γονιδίων αλλά αντιπροσωπευτικά ids για το κάθε γονίδιο), την στάθμιση που έχει με βάση τις επτά έρευνες και την κλάση στην οποία ανήκει, και το αρχείο όπου μέσα σε αυτό δίνονται τα σχετικά βάρη, που είναι επτά τιμές όπου η κάθε τιμή αντιπροσωπεύει την βαρύτητα που δίνεται σε κάθε έρευνα.

2.7.1 Δημιουργία αντιπροσωπευτικής κλάσεις με βάση τα σταθμισμένα βάρη που έχει το κάθε γονίδιο.

Στην Εικόνα 2.7.1 παρουσιάζεται τυπικό δείγμα από τα δεδομένα που αργότερα θα εισάγουμε μαζί με το χαρακτηριστικό id του κάθε γονιδίου στον αλγόριθμο μας μέσω του αρχείου «data.txt» Οι πρώτες επτά στήλες που έχει το παράδειγμα μας αντιπροσωπεύουν τις επτά διαφορετικές μελέτες. Η τελευταία στήλη αντιπροσωπεύει την κλάση. Όπως έχουμε αναφέρει και πιο πάνω η κάθε γραμμή αντιπροσωπεύει το κάθε γονίδιο και τις διάφορες τιμές που παίρνει από κάθε διαφορετική μελέτη . Αν το γονίδιο βρεθεί να έχει κάποια συσχέτιση με την ασθένεια που αναζητούμε τότε παίρνει μία τιμή συσχέτισης. Αν όμως δεν έχει βρεθεί να έχει κάποια συσχέτιση τότε παίρνει τιμή ποινής

(Faul) που είναι -99. Για παράδειγμα το πρώτο γονίδιο της πρώτης γραμμής δεν βρέθηκε να έχει συσχέτιση με την ασθένεια που αναζητούμε στην 3 μελέτη.

Εικόνα 2.7.1: Τυπικό δείγμα από δεδομένα που επεξεργάζονται για να εισαχθούν στον αλγόριθμο Metradisc. Στο παρών δείγμα επεξηγείται ο τρόπος μέσα από τον οποίο με βάση τις σταθμίσεις που έχει το κάθε γονίδιο(επτά πρώτες στήλες) βρίσκουμε την κλάση στην οποία ανήκει.

3406.0	7995.0	-99.0	194.5	1663.5	1237.5	487.0	=	3.0
-99.0	6010.5	5268.0	-99.0	-99.0	3746.0	6411.5	=	40.0
5881.0	6670.0	-99.0	-99.0	1352.5	-99.0	8668.0	=	57.0
3989.0	6910.0	3562.0	3111.0	718.5	6227.0	5245.5	=	132.0
-99.0	2751.5	-99.0	-99.0	-99.0	-99.0	8038.5	=	113.0
5739.0	2523.5	-99.0	-99.0	126.5	-99.0	174.0	=	57.0
-99.0	-99.0	3719.0	2382.5	-99.0	4582.5	2437.5	=	33.0
-99.0	-99.0	-99.0	-99.0	-99.0	-99.0	7237.5	=	125.0
-99.0	2537.5	-99.0	-99.0	-99.0	-99.0	-99.0	=	130.0
5739.0	8532.0	5130.0	4556.5	1283.5	2526.0	8365.5	=	132.0
4994.0	471.0	1660.5	1708.0	-99.0	3938.0	3993.5	=	5.0
-99.0	3732.5	-99.0	-99.0	-99.0	-99.0	-99.0	=	130.0
-99.0	-99.0	2196.5	194.5	-99.0	164.0	555.0	=	33.0
3553.0	7458.5	4214.5	5913.5	-99.0	6364.0	6040.0	=	5.0
-99.0	1339.5	5144.5	5770.5	-99.0	2495.0	3032.0	=	12.0
-99.0	-99.0	6067.0	2220.5	-99.0	3995.0	7067.5	=	33.0
427.0	6299.0	-99.0	-99.0	1653.5	-99.0	2007.0	=	57.0
-99.0	4866.5	-99.0	-99.0	-99.0	-99.0	6479.0	=	113.0
4595.0	-99.0	-99.0	-99.0	3350.0	-99.0	-99.0	=	120.0
-99.0	7228.5	-99.0	-99.0	-99.0	-99.0	3177.5	=	113.0
-99.0	8403.5	-99.0	-99.0	-99.0	-99.0	-99.0	=	130.0
1605.0	-99.0	586.5	-99.0	4925.5	3372.0	5911.5	=	16.0
6228.0	-99.0	-99.0	-99.0	2060.0	-99.0	7067.5	=	88.0
-99.0	8967.5	3149.5	3364.5	-99.0	6227.0	-99.0	=	44.0
4723.0	7867.0	5528.5	5656.0	3203.0	4707.5	6857.0	=	132.0
2562.0	-99.0	-99.0	-99.0	4173.5	-99.0	-99.0	=	120.0
-99.0	471.0	5709.0	5629.5	3768.5	1523.5	4119.5	=	1.0
-99.0	-99.0	-99.0	-99.0	-99.0	-99.0	4919.0	=	125.0
-99.0	2965.5	4709.5	2083.0	-99.0	859.0	-99.0	=	44.0
-99.0	-99.0	4821.0	2988.0	-99.0	1033.0	-99.0	=	75.0

Για να δημιουργηθούν οι κλάσεις, με βάση τα δεδομένα που έχουμε, δημιουργούμε ένα πίνακα με τιμές 0 και 1, όπου παρουσιάζεται η τιμή -99 (ποινή) παίρνει την τιμή 1 και όπου άλλη τιμή παίρνει 0. Για παράδειγμα τα δεδομένα του πρώτου γονιδίου μεταφράζονται ως:

0 0 1 0 0 0 0 = 3 κλάση

Με βάση την κωδικοποίηση τους σε 0 και 1 τότε βλέπουμε σε πια κλάση πρέπει να τα ορίσουμε το κάθε γονίδιο με βάση τον πίνακα 2.7.1. που έχει όλες τις

αντιπροσωπευτικές κλάσεις που μπορεί να υπάρξουν. Ο καθορισμός του αριθμού των κλάσεων K δίνεται από τον ακόλουθο μαθηματικό τύπο:

$$K = 2^S + B$$

όπου s αντιπροσωπεύει τον αριθμό των ερευνών που χρησιμοποιούμε και B ο αριθμός των κενών περιοχών που δημιουργούνται κατά το γέμισμα του πίνακα. Με βάση τον πιο πάνω μαθηματικό τύπο στην παρούσα εργασία χρησιμοποιήθηκαν 132 κλάσεις. (Εικόνα 2.7.2)

Εικόνα 2.7.2: Ενδεικτικό ολοκληρωμένο απόσπασμα από τα δεδομένα που εισάγονται στον αλγόριθμο σε μορφή αρχείου «data.txt» όπως βρέθηκαν για τον Καρκίνο του μαστού.

1	39366	-99	-99	-99	-99	-99	-99	1
2	39367	-99	-99	-99	-99	-99	-99	1
3	39368	-99	-99	-99	-99	-99	-99	1
4	39369	39□10	-□9	-□9	-99	-99	39161	35
5	39370	-99	-99	-99	-99	-99	-99	1
6	39371	39613	39494	-99	39836	39927	39109	128
7	39372	-99	-99	-99	-99	-99	-99	1
8	39373	39463	39459	39807	39835	39925	39108	1
9	39374	-99	-99	-99	39842	-99	39132	44
10	39375	-99	39856	-99	39891	-99	-99	37
11	39376	-99	39847	-99	39864	-99	39218	81
12	39377	-99	-99	-99	-99	-99	-99	1
13	39378	-99	-99	-99	39870	-99	39443	44
14	39379	-99	39622	-99	39858	-99	39131	81
15	39380	-99	39693	-99	-99	-99	-99	10
16	39381	-99	39926	-99	39863	-99	39237	81
17	39382	-99	-99	-99	39916	-99	-99	12
18	39383	-99	-99	-99	39859	-99	39720	44
19	39384	39580	-99	-99	39848	-99	39233	75
20	39385	-99	-99	-99	-99	-99	-99	1
21	39386	-99	39639	-99	39846	-99	39142	81
22	39387	-99	39781	-99	39854	-99	39121	81
23	39388	-99	-99	-99	-99	-99	-99	1
24	39389	39538	39483	39868	39895	39901	39111	1
25	39390	39450	39564	39806	39881	39903	39110	1

Πίνακας 2.7.1: Όλες οι αντιπροσωπευτικές κλάσεις που δημιουργούνται για ανάλυση των δεδομένων που εισάγουμε στο Metradisc.

1	1	0	0	0	0	0	0	48	0	1	1	0	0	1	0	95	0	1	0	1	0	1	1
2	0	1	0	0	0	0	0	49	0	1	1	0	0	0	1	96	0	1	0	0	1	1	1
3	0	0	1	0	0	0	0	50	0	1	0	1	1	0	0	97	0	0	1	1	1	1	0
4	0	0	0	1	0	0	0	51	0	1	0	1	0	1	0	98	0	0	1	1	1	0	1
5	0	0	0	0	1	0	0	52	0	1	0	1	0	0	1	99	0	0	1	1	0	1	1
6	0	0	0	0	0	1	0	53	0	1	0	0	1	1	0	100	0	0	1	0	1	1	1
7	0	0	0	0	0	0	1	54	0	1	0	0	1	0	1	101	0	0	0	1	1	1	1
8	Κενή περιοχή							55	0	1	0	0	0	1	1	102	Κενή περιοχή						
9	1	1	0	0	0	0	0	56	0	0	1	1	1	0	0	103	1	1	1	1	1	0	0
10	1	0	1	0	0	0	0	57	0	0	1	1	0	1	0	104	1	1	1	1	0	1	0
11	1	0	0	1	0	0	0	58	0	0	1	1	0	0	1	105	1	1	1	1	0	0	1
12	1	0	0	0	1	0	0	59	0	0	1	0	1	1	0	106	1	1	1	0	1	1	0
13	1	0	0	0	0	1	0	60	0	0	1	0	1	0	1	107	1	1	1	0	1	0	1
14	1	0	0	0	0	0	1	61	0	0	1	0	0	1	1	108	1	1	1	0	0	1	1
15	0	1	1	0	0	0	0	62	0	0	0	1	1	1	0	109	1	1	0	1	1	1	0
16	0	1	0	1	0	0	0	63	0	0	0	1	1	0	1	110	1	1	0	1	1	0	1
17	0	1	0	0	1	0	0	64	0	0	0	1	0	1	1	111	1	0	0	1	0	1	1
18	0	1	0	0	0	1	0	65	0	0	0	0	1	1	1	112	1	0	0	0	1	1	1
19	0	1	0	0	0	0	1	66	Κενή περιοχή							113	1	0	1	1	1	1	0
20	0	0	1	1	0	0	0	67	1	1	1	1	0	0	0	114	1	0	1	1	1	0	1
21	0	0	1	0	1	0	0	68	1	1	1	0	1	0	0	115	1	0	1	1	0	1	1
22	0	0	1	0	0	1	0	69	1	1	1	0	0	1	0	116	1	0	1	0	1	1	1
23	0	0	1	0	0	0	1	70	1	1	1	0	0	0	1	117	1	0	0	1	1	1	1
24	0	0	0	1	1	0	0	71	1	1	0	1	1	0	0	118	0	1	1	1	1	1	0
25	0	0	0	1	0	1	0	72	1	1	0	1	0	1	0	119	0	1	1	1	1	0	1
26	0	0	0	1	0	0	1	73	1	1	0	1	0	0	1	120	0	1	1	1	0	1	1
27	0	0	0	0	1	1	0	74	1	1	0	0	1	1	0	121	0	1	1	0	1	1	1
28	0	0	0	0	1	0	1	75	1	1	0	0	1	0	1	122	0	1	0	1	1	1	1
29	0	0	0	0	0	1	1	76	1	1	0	0	0	1	1	123	0	0	1	1	1	1	1
30	Κενή περιοχή							77	1	0	1	1	1	0	0	124	Κενή περιοχή						
31	1	1	1	0	0	0	0	78	1	0	1	1	0	1	0	125	1	1	1	1	1	1	0
32	1	1	0	1	0	0	0	79	1	0	1	1	0	0	1	126	1	1	1	1	1	0	1
33	1	1	0	0	1	0	0	80	1	0	1	0	1	1	0	127	1	1	1	1	0	1	1
34	1	1	0	0	0	1	0	81	1	0	1	0	1	0	1	128	1	1	1	0	1	1	1
35	1	1	0	0	0	0	1	82	1	0	1	0	0	1	1	129	1	1	0	1	1	1	1
36	1	0	1	1	0	0	0	83	1	0	0	1	1	1	0	130	1	0	1	1	1	1	1
37	1	0	1	0	1	0	0	84	1	0	0	1	1	0	1	131	0	1	1	1	1	1	1
38	1	0	1	0	0	1	0	85	1	0	0	1	0	1	1	132	0	0	0	0	0	0	0
39	1	0	1	0	0	0	1	86	1	0	0	0	1	1	1								
40	1	0	0	1	1	0	0	87	0	1	1	1	1	0	0								
41	1	0	0	1	0	1	0	88	0	1	1	1	0	1	0								
42	1	0	0	1	0	0	1	89	0	1	1	1	0	0	1								
43	1	0	0	0	1	1	0	90	0	1	1	0	1	1	0								
44	1	0	0	0	1	0	1	91	0	1	1	0	1	0	1								
45	1	0	0	0	0	1	1	92	0	1	1	0	0	1	1								
46	0	1	1	1	0	0	0	93	0	1	0	1	1	1	0								
47	0	1	1	0	1	0	0	94	0	1	0	1	1	0	1								

2.7.2 Υπολογισμός αντιπροσωπευτικού βάρους για κάθε έρευνα.

Όπως έχουμε αναφέρει και πιο πάνω το δεύτερο όρισμα που παίρνει ο αλγόριθμος Metradisc είναι τα αντιπροσωπευτικά βάρη που δίνονται μέσο του αρχείου «weights.txt» στην κάθε έρευνα. Ο σκοπός που γίνεται αυτή η διαδικασία είναι για να υπάρχει μία ισότητα μεταξύ των ερευνών που είναι κατά πολύ μεγαλύτερες ή κατά πολύ μικρότερες από κάποιες άλλες.

Το αντιπροσωπευτικό βάρος Β που πρέπει να δίνεται για κάθε έρευνα δίνεται από τον ακόλουθο τύπου:

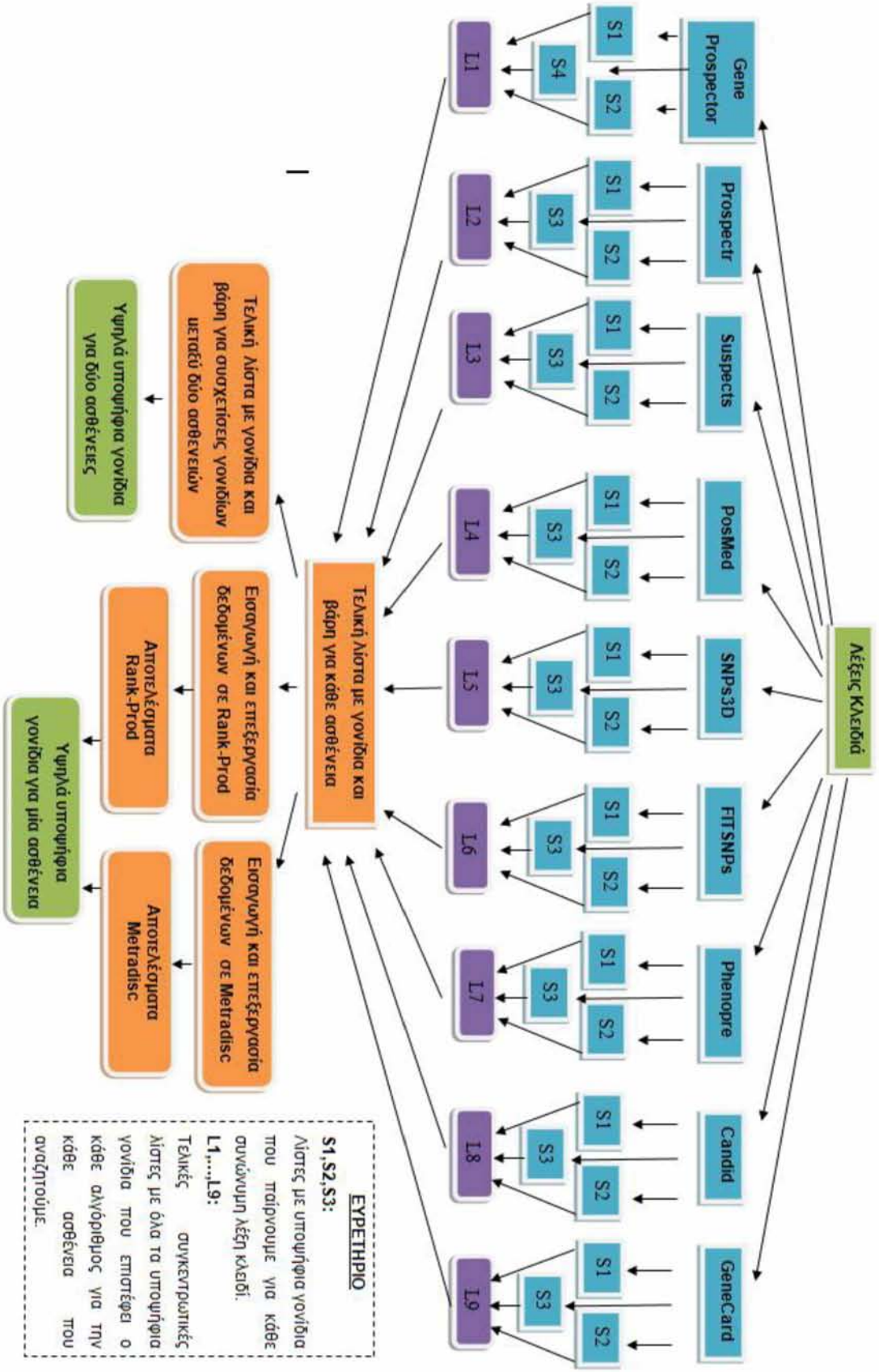
$$B = G_t / G_s$$

Όπου G_tείναι το σύνολο όλων των γονιδίων που έχουμε ανακτήσει για μία ασθένεια και G_sτο σύνολο γονιδίων που έχει η κάθε s μελέτη ξεχωριστά.

Στην ακραία περίπτωση που και όλες οι μελέτες μας περιέχουν ίσο αριθμό από γονίδια η κάθε μία ξεχωριστά τότε πολύ απλά τα βάρη μας είναι για όλες τα ίδια και συνεπώς θα έχουν και οι επτά έρευνες αντιπροσωπευτικό βάρος την μονάδα. Δηλαδή θα εισάγουμε μέσω τους αρχείου «weights.txt» τις τιμές :

1.0 1.0 1.0 1.0 1.0 1.0 1.0

2.8 Αναλυτικό διάγραμμα ροής διαδικασιών .



ΚΕΦΑΛΑΙΟ 3: ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Αποτελέσματα MetraDisc.

Ο αλγόριθμος στατιστικής ανάλυσης Metradisc έχει εφαρμοστεί σε δεδομένα από συνδυασμό των έπτα διαφορετικών μελετών(αλγόριθμοι αναζήτησης υποψήφιων γονιδίων που σχετίζονται με ασθένειες). Τα σταθμισμένα δεδομένα παράγωνται με βάση την ιεράρχηση που έπαιρνε το κάθε γονίδιο στον κάθε αλγόριθμο αναζήτησης από τον οποίο προήλθε. Οι ποινές (faul) δίνονταν στα γονίδια που δεν βρέθηκαν σε κάποιο από τους επτά αλγόριθμους που χρησιμοποιήσαμε.

Η στατιστική ανάλυση του αλγόριθμου εφαρμόστηκε και στα έξι σύνολα δεδομένων που είχαμε για τις ασθένειες Καρκίνο του μαστου, Διαβήτη τύπου Ι, Διαβήτη τύπου ΙΙ, Υπέρταση, Σκλήρυνση κατά πλάκα και Παχυσαρκία. Τα αποτελέσματα παρουσίαζονται στους ακόλουθους πίνακες 3.1.1-3.1.6 .

Οι μεταθέσεις ελέγχου χρησιμοποιώντας μηδενική κατανομή ήταν 100 προσομοιώσεις για τις R* και Q*μετρήσεις. Έτσι, κάθε γονίδιο έχει ελεγχθεί σε μια μηδενική κατανομή 100 g μετρήσεων, όπου g είναι ο αριθμός των γονιδίων με την ίδια κατηγορία πληροφοριών. Για παράδειγμα, η ελάχιστη τιμή g ήταν μία, επομένως, όλα τα γονίδια έχουν δοκιμαστεί με βάση τις ελάχιστες 100 μεταθέσεις σε μηδενική κατανομή.

Κανένα από αυτά που έχουν δοκιμαστεί λιγότερες φορές έχουν αριστερό ή δεξιό μονόπλευρο p-value για το R*. Αυτό το R* δεν μπορεί να διαφοροποιηθεί με ακρίβεια από το μηδέν (Λόγω της αραιής προσομοίωσης με τιμές στη μηδενική κατανομή). Για την εύρεση της ετερογένειας Q ένας μεγαλύτερος αριθμός από προσομοιώσεις είναι απαραίτητος, δεδομένου ότι μόνο ένα υποσύνολο των προσομοιωμένων τιμών τελικά χρησιμοποιείται για την κατασκευή της μηδενικής κατανομής για ένα συγκεκριμένο γονίδιο. Τα αποτελέσματα βασίζονται σε 100.000 g προσομοιώσεις. Η διαδικασία είναι υπολογιστικά χρονοβόρα. Για τα γονίδια με πολύ ακραίες τιμές στο R* είναι δύσκολο να δημιουργηθεί ένα αρκετά μεγάλο δείγμα από επιλεγμένα προσομοιωμένα γονίδια με παρόμοιες προσομοιώσεις στο R* για να πάρει πολύ συγκεκριμένες τιμές το p-value για την Q*.

Πίνακας 3.1.1: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με τον Καρκίνο του μαστού.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
231	AR	39596	39480	-99	39866	39892	39917	39211	129	6.56E+05	8.10E-01	1.90E-01	8.10E-01	1.90E-01	3.97E+12	2.30E-01	7.70E-01
64	RAD51	39429	39748	39501	-99	39908	39910	39463	128	5.93E+05	8.15E-01	1.85E-01	8.15E-01	1.85E-01	4.20E+12	5.95E-01	4.05E-01
387	TP53	39752	39431	39475	-99	39869	39921	39112	128	5.92E+05	9.50E-01	5.00E-02	9.50E-01	5.00E-02	4.20E+12	7.50E-02	9.25E-01
334	CHEK2	39699	39698	39519	-99	39887	39920	39133	128	5.93E+05	7.65E-01	2.35E-01	7.65E-01	2.35E-01	4.20E+12	2.05E-01	7.95E-01
6	BRCA2	39371	39613	39494	-99	39836	39927	39109	128	5.92E+05	9.83E-01	1.75E-02	9.83E-01	1.75E-02	4.20E+12	0	1
39	CYP19A1	39404	-99	39566	-99	39910	39902	39147	116	6.85E+05	9.90E-01	1.00E-02	9.90E-01	1.00E-02	3.94E+12	7.70E-01	2.30E-01
45	BARD1	39410	-99	39495	39841	-99	39924	39205	115	6.49E+05	9.40E-01	6.00E-02	9.40E-01	6.00E-02	3.96E+12	0	1
409	AURKA	39774	-99	39460	39917	39913	-99	39197	114	3.09E+05	3.90E-01	6.10E-01	3.90E-01	6.10E-01	3.15E+11	2.00E-02	9.80E-01
70	CDH1	39435	39536	39484	-99	-99	39909	39113	108	5.71E+05	1	0	1	0	4.18E+12	4.20E-01	5.80E-01
202	NCOA3	39567	39658	39462	-99	39840	-99	39158	107	2.31E+05	9.90E-01	1.00E-02	9.90E-01	1.00E-02	2.73E+11	7.50E-01	2.50E-01
138	MDM2	39503	39629	39465	-99	39912	-99	39208	107	2.31E+05	8.65E-01	1.35E-01	8.65E-01	1.35E-01	2.74E+11	4.50E-02	9.55E-01
108	CCND1	39473	39720	39464	-99	39923	-99	39153	107	2.31E+05	7.93E-01	2.08E-01	7.93E-01	2.08E-01	2.74E+11	7.50E-03	9.93E-01
26	ESR2	39391	39758	39669	-99	39915	-99	39200	107	2.31E+05	6.48E-01	3.53E-01	6.48E-01	3.53E-01	2.74E+11	6.00E-02	9.40E-01
48	MAPK1	39413	39447	39911	39852	-99	-99	39259	105	1.95E+05	8.21E-01	1.79E-01	8.21E-01	1.79E-01	1.32E+11	3.39E-01	6.61E-01
154	STAT3	39519	39442	39587	39839	-99	-99	39337	105	1.95E+05	9.65E-01	3.50E-02	9.65E-01	3.50E-02	1.32E+11	3.85E-01	6.15E-01
471	RAC1	39836	39458	39700	39920	-99	-99	39588	105	1.95E+05	1.31E-01	8.69E-01	1.31E-01	8.69E-01	1.33E+11	1.41E-01	8.59E-01
44	CASP3	39409	39462	39609	39860	-99	-99	39244	105	1.95E+05	9.69E-01	3.14E-02	9.69E-01	3.14E-02	1.32E+11	2.19E-01	7.81E-01
42	BCL2	39407	39510	39605	39844	-99	-99	39144	105	1.95E+05	9.84E-01	1.64E-02	9.84E-01	1.64E-02	1.32E+11	2.76E-01	7.24E-01
46	MAPK3	39411	39453	39923	39836	-99	-99	39399	105	1.95E+05	7.89E-01	2.11E-01	7.89E-01	2.11E-01	1.32E+11	5.44E-01	4.56E-01
251	SRC	39616	39428	39610	39832	-99	-99	39134	105	1.95E+05	9.72E-01	2.79E-02	9.72E-01	2.79E-02	1.32E+11	4.00E-01	6.01E-01
33	KDR	39398	39731	39801	39909	-99	-99	39367	105	1.95E+05	3.09E-01	6.91E-01	3.09E-01	6.91E-01	1.33E+11	1.05E-01	8.95E-01
142	CDKN1B	39507	39679	39637	39916	-99	-99	39332	105	1.95E+05	4.53E-01	5.47E-01	4.53E-01	5.47E-01	1.33E+11	4.71E-02	9.53E-01

137	CDKN1A	39502	39517	39478	39837	-99	-99	39182	105	1.95E+05	9.91E-01	8.57E-03	9.91E-01	8.57E-03	1.32E+11	3.31E-01	6.69E-01
79	MYC	39444	39518	39468	39816	-99	-99	39186	105	1.95E+05	9.98E-01	2.14E-03	9.98E-01	2.14E-03	1.32E+11	4.66E-01	5.34E-01
145	IGF1R	39510	39588	39600	39846	-99	-99	39236	105	1.95E+05	9.17E-01	8.29E-02	9.17E-01	8.29E-02	1.32E+11	3.54E-01	6.46E-01
58	PIK3R1	39423	39430	39711	39808	-99	-99	39564	105	1.95E+05	9.50E-01	5.00E-02	9.50E-01	5.00E-02	1.32E+11	8.21E-01	1.79E-01
140	PCNA	39505	39617	39864	39857	-99	-99	39168	105	1.95E+05	6.46E-01	3.54E-01	6.46E-01	3.54E-01	1.32E+11	3.44E-01	6.56E-01
943	XRCC3	-99	-99	39531	-99	39897	39911	39178	100	8.28E+05	9.40E-01	6.00E-02	9.40E-01	6.00E-02	3.53E+12	2.55E-01	7.45E-01
948	NAT2	-99	-99	39537	-99	39904	39918	39152	100	8.28E+05	8.00E-01	2.00E-01	8.00E-01	2.00E-01	3.53E+12	4.50E-02	9.55E-01
916	PHB	-99	-99	39489	39869	-99	39914	39240	99	7.83E+05	9.00E-01	1.00E-01	9.00E-01	1.00E-01	3.60E+12	8.00E-02	9.20E-01
757	PTEN	-99	39745	39461	-99	-99	39912	39116	92	6.86E+05	9.50E-01	5.00E-02	9.50E-01	5.00E-02	3.92E+12	1.70E-01	8.30E-01
596	JUN	-99	39512	39691	39889	-99	-99	39271	89	2.15E+05	7.28E-01	2.73E-01	7.28E-01	2.73E-01	1.24E+11	1.25E-01	8.75E-01
613	E2F1	-99	39539	39927	39831	-99	-99	39474	89	2.16E+05	4.98E-01	5.03E-01	4.98E-01	5.03E-01	1.24E+11	8.35E-01	1.65E-01
865	CCNA2	-99	39882	39830	39895	-99	-99	39280	89	2.16E+05	9.75E-02	9.03E-01	9.75E-02	9.03E-01	1.24E+11	3.55E-01	6.45E-01
646	HDAC1	-99	39583	39805	39891	-99	-99	39691	89	215744	1.68E-01	8.33E-01	1.68E-01	8.33E-01	1.24E+11	4.23E-01	5.78E-01
269	BRIP1	39634	-99	39530	-99	-99	39926	39196	82	6.83E+05	7.90E-01	2.10E-01	7.90E-01	2.10E-01	3.94E+12	4.00E-02	9.60E-01
40	SHBG	39405	-99	39535	-99	-99	39916	39378	82	6.82E+05	9.65E-01	3.50E-02	9.65E-01	3.50E-02	3.94E+12	3.10E-01	6.90E-01
31	VEGFA	39396	-99	39540	-99	39902	-99	39138	81	2.56E+05	9.53E-01	4.71E-02	9.53E-01	4.71E-02	2.61E+11	2.71E-02	9.73E-01
27	PGR	39392	-99	39630	-99	39899	-99	39122	81	2.56E+05	9.21E-01	7.88E-02	9.21E-01	7.88E-02	2.61E+11	4.88E-02	9.51E-01
11	BCAS1	39376	-99	39847	-99	39864	-99	39218	81	2.56E+05	8.66E-01	1.34E-01	8.66E-01	1.34E-01	2.60E+11	5.90E-01	4.10E-01
113	IGF1	39478	-99	39799	-99	39900	-99	39204	81	2.56E+05	6.39E-01	3.61E-01	6.39E-01	3.61E-01	2.60E+11	1.74E-01	8.26E-01
260	PTGS2	39625	-99	39574	-99	39893	-99	39180	81	2.56E+05	8.20E-01	1.80E-01	8.20E-01	1.80E-01	2.60E+11	1.95E-01	8.05E-01
111	IGFBP3	39476	-99	39488	-99	39917	-99	39216	81	2.56E+05	8.75E-01	1.25E-01	8.75E-01	1.25E-01	2.61E+11	9.41E-03	9.91E-01
119	MTHFR	39484	-99	39755	-99	39886	-99	39160	81	2.56E+05	8.04E-01	1.96E-01	8.04E-01	1.96E-01	2.60E+11	2.94E-01	7.06E-01
16	BCAS3	39381	-99	39926	-99	39863	-99	39237	81	2.56E+05	7.92E-01	2.08E-01	7.92E-01	2.08E-01	2.60E+11	6.54E-01	3.46E-01
14	BRMS1	39379	-99	39622	-99	39858	-99	39131	81	2.56E+05	9.89E-01	1.06E-02	9.89E-01	1.06E-02	2.60E+11	4.76E-01	5.24E-01
378	RHOBTB2	39743	-99	39727	-99	39862	-99	39530	81	2.57E+05	4.25E-01	5.75E-01	4.25E-01	5.75E-01	2.59E+11	8.53E-01	1.47E-01
358	CYP1A1	39723	-99	39511	-99	39882	-99	39129	81	2.56E+05	8.73E-01	1.27E-01	8.73E-01	1.27E-01	2.60E+11	3.22E-01	6.78E-01
226	CYP17A1	39591	-99	39541	-99	39890	-99	39333	81	2.56E+05	8.09E-01	1.91E-01	8.09E-01	1.91E-01	2.60E+11	2.74E-01	7.26E-01

Πίνακας 3.1.2: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με τον Διαβήτη τύπου Ι.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
63	CTLA4	39237	39823	39846	-99	39778	39924	39154	128	1.12E+06	8.80E-01	1.20E-01	8.80E-01	1.20E-01	2.07E+13	1.50E-01	8.50E-01
1350	SUMO4	-99	-99	39852	39903	39807	39914	39192	123	1.39E+06	9.20E-01	8.00E-02	9.20E-01	8.00E-02	1.86E+13	8.30E-01	1.70E-01
181	PTPN22	39355	-99	39830	-99	39780	39926	39145	116	1.32E+06	9.38E-01	6.25E-02	9.38E-01	6.25E-02	1.95E+13	7.50E-03	9.93E-01
588	TAP2	39762	-99	39901	-99	39773	39916	39166	116	1.32E+06	8.13E-01	1.88E-01	8.13E-01	1.88E-01	1.95E+13	7.70E-01	2.30E-01
26	IL6	39200	-99	39868	-99	39783	39923	39133	116	1.32E+06	9.55E-01	4.50E-02	9.55E-01	4.50E-02	1.95E+13	5.50E-02	9.45E-01
77	VDR	39251	-99	39875	-99	39775	39922	39175	116	1.32E+06	9.40E-01	6.00E-02	9.40E-01	6.00E-02	1.95E+13	1.10E-01	8.90E-01
30	IFNG	39204	-99	-99	-99	39782	39925	39223	112	1.48E+06	9.33E-01	6.67E-02	9.33E-01	6.67E-02	1.89E+13	1.67E-02	9.83E-01
93	IL18	39267	-99	-99	-99	39784	39927	39324	112	1.48E+06	7.73E-01	2.27E-01	7.73E-01	2.27E-01	1.89E+13	3.33E-03	9.97E-01
40	ICAM1	39214	-99	-99	-99	39774	39917	39262	112	1.48E+06	9.90E-01	1.00E-02	9.90E-01	1.00E-02	1.89E+13	4.47E-01	5.53E-01
202	PTPN1	39376	39294	-99	39925	39910	-99	39453	110	2.60E+05	4.00E-01	6.00E-01	4.00E-01	6.00E-01	1.85E+11	4.00E-02	9.60E-01
240	GCK	39414	-99	-99	39806	39808	-99	39128	84	2.97E+05	9.53E-01	4.75E-02	9.53E-01	4.75E-02	1.56E+11	2.93E-01	7.10E-01
241	GCKR	39415	-99	-99	39814	39813	-99	39635	84	2.97E+05	6.38E-01	3.63E-01	6.38E-01	3.63E-01	1.55E+11	6.50E-01	3.50E-01
248	PPARGC1A	39422	-99	-99	39776	39820	-99	39408	84	2.97E+05	8.98E-01	1.03E-01	8.98E-01	1.03E-01	1.56E+11	5.90E-01	4.10E-01
513	ADRB3	39687	-99	-99	39812	39810	-99	39248	84	2.97E+05	7.28E-01	2.73E-01	7.28E-01	2.73E-01	1.56E+11	5.80E-01	4.20E-01
188	AVPR2	39362	-99	39845	-99	-99	39920	39181	82	1.52E+06	9.90E-01	1.00E-02	9.90E-01	1.00E-02	1.86E+13	2.30E-01	7.70E-01
162	VEGFA	39336	-99	39873	-99	39846	-99	39142	81	3.36E+05	8.58E-01	1.42E-01	8.58E-01	1.42E-01	2.35E+11	7.82E-02	9.22E-01
34	LTA	39208	-99	39870	-99	39892	-99	39306	81	3.36E+05	6.79E-01	3.21E-01	6.79E-01	3.21E-01	2.36E+11	4.45E-02	9.55E-01
444	LPL	39618	-99	39918	-99	39828	-99	39150	81	3.36E+05	4.37E-01	5.63E-01	4.37E-01	5.63E-01	2.35E+11	6.64E-02	9.34E-01
227	HNF1A	39401	-99	39833	-99	39814	-99	39147	81	3.36E+05	9.65E-01	3.45E-02	9.65E-01	3.45E-02	2.35E+11	3.99E-01	6.01E-01
230	PAX4	39404	-99	39829	-99	39927	-99	39198	81	3.36E+05	6.71E-01	3.29E-01	6.71E-01	3.29E-01	2.35E+11	1.79E-01	8.21E-01
597	WFS1	39771	-99	39842	-99	39829	-99	39177	81	3.36E+05	7.14E-01	2.87E-01	7.14E-01	2.87E-01	2.34E+11	6.54E-01	3.46E-01
185	PPARG	39359	-99	39879	-99	39785	-99	39135	81	3.36E+05	9.45E-01	5.55E-02	9.45E-01	5.55E-02	2.35E+11	1.52E-01	8.48E-01
24	TNF	39198	-99	39867	-99	39795	-99	39189	81	3.36E+05	9.74E-01	2.64E-02	9.74E-01	2.64E-02	2.35E+11	1.30E-01	8.70E-01
66	IL2RA	39240	-99	39836	-99	39811	-99	39362	81	3.36E+05	9.58E-01	4.18E-02	9.58E-01	4.18E-02	2.35E+11	4.28E-01	5.72E-01
158	NEUROD1	39332	-99	39865	-99	39831	-99	39185	81	3.36E+05	8.98E-01	1.02E-01	8.98E-01	1.02E-01	2.35E+11	1.44E-01	8.56E-01
121	APOC3	39295	-99	39869	-99	39848	-99	39379	81	3.36E+05	7.43E-01	2.57E-01	7.43E-01	2.57E-01	2.35E+11	1.74E-01	8.26E-01
146	FGF2	39320	39607	-99	-99	39912	-99	39363	75	2.02E+05	4.70E-01	5.30E-01	4.70E-01	5.30E-01	1.16E+11	1.00E-01	9.00E-01
64	PTPRC	39238	39333	-99	39837	-99	-99	39331	73	2.01E+05	9.00E-01	1.00E-01	9.00E-01	1.00E-01	1.14E+11	1.40E-01	8.60E-01
86	AKT1	39260	39236	-99	39771	-99	-99	39396	73	2.01E+05	9.83E-01	1.67E-02	9.83E-01	1.67E-02	1.14E+11	4.57E-01	5.43E-01
5	IGF1R	39179	39414	-99	39773	-99	-99	39274	73	200628	9.83E-01	1.67E-02	9.83E-01	1.67E-02	1.14E+11	4.37E-01	5.63E-01

258	FASLG	39432	39479	39903	-99	-99	-99	39425	70	2.40E+05	4.95E-01	5.05E-01	4.95E-01	5.05E-01	2.23E+11	1.25E-01	8.75E-01
32	TNFRSF1A	39206	39347	39878	-99	-99	-99	39384	70	2.40E+05	9.00E-01	1.00E-01	9.00E-01	1.00E-01	2.23E+11	1.40E-01	8.60E-01
1427	INS	-99	-99	-99	39826	39796	-99	39129	63	3.62E+05	9.70E-01	3.00E-02	9.70E-01	3.00E-02	1.06E+11	1.50E-01	8.50E-01
1347	CLEC16A	-99	-99	39847	-99	39818	-99	39787	60	414186	4.70E-01	5.30E-01	4.70E-01	5.30E-01	1.62E+11	9.17E-01	8.33E-02
1358	MICA	-99	-99	39872	-99	39804	-99	39335	60	4.14E+05	7.90E-01	2.10E-01	7.90E-01	2.10E-01	1.62E+11	2.83E-01	7.17E-01
1349	MBL2	-99	-99	39851	-99	39926	-99	39454	60	4.14E+05	2.67E-01	7.33E-01	2.67E-01	7.33E-01	1.62E+11	3.63E-01	6.37E-01
989	CCR5	-99	39535	-99	-99	39924	-99	39302	54	2.35E+05	3.65E-01	6.35E-01	3.65E-01	6.35E-01	1.03E+11	2.83E-02	9.72E-01
1115	LEPR	-99	39686	-99	-99	39830	-99	39352	54	2.35E+05	5.77E-01	4.23E-01	5.77E-01	4.23E-01	1.03E+11	5.43E-01	4.57E-01
979	ERBB3	-99	39524	-99	-99	39809	-99	39739	54	2.35E+05	4.97E-01	5.03E-01	4.97E-01	5.03E-01	1.02E+11	8.13E-01	1.87E-01
1003	AGTR1	-99	39555	-99	-99	39833	-99	39158	54	234680	8.43E-01	1.57E-01	8.43E-01	1.57E-01	1.03E+11	3.20E-01	6.80E-01
939	PTPN2	-99	39481	-99	-99	39844	-99	39838	54	2.35E+05	2.97E-01	7.03E-01	2.97E-01	7.03E-01	1.03E+11	6.43E-01	3.57E-01
1036	ADRB2	-99	39591	-99	-99	39861	-99	39186	54	2.35E+05	6.80E-01	3.20E-01	6.80E-01	3.20E-01	1.03E+11	2.27E-01	7.73E-01
967	AKT2	-99	39511	-99	39922	-99	-99	39601	52	2.34E+05	1.70E-01	8.30E-01	1.70E-01	8.30E-01	1.02E+11	1.43E-01	8.57E-01
866	GAPDH	-99	39390	-99	39803	-99	-99	39520	52	2.34E+05	8.17E-01	1.83E-01	8.17E-01	1.83E-01	1.01E+11	6.63E-01	3.37E-01
1107	PTPRF	-99	39676	-99	39836	-99	-99	39487	52	2.34E+05	4.17E-01	5.83E-01	4.17E-01	5.83E-01	1.01E+11	6.37E-01	3.63E-01
793	ESR1	-99	39288	-99	39822	39919	-99	-99	50	3.67E+05	4.70E-01	5.40E-01	4.70E-01	5.40E-01	9.87E+10	1.30E-01	8.80E-01
982	NPHS1	-99	39528	39880	-99	-99	-99	39163	49	2.86E+05	6.73E-01	3.27E-01	6.73E-01	3.27E-01	1.98E+11	1.17E-01	8.83E-01
1018	PRKAR1A	-99	39570	39810	-99	-99	-99	39382	49	2.85E+05	8.40E-01	1.60E-01	8.40E-01	1.60E-01	1.97E+11	7.97E-01	2.03E-01
933	ACP1	-99	39471	39812	-99	-99	-99	39317	49	2.85E+05	9.20E-01	8.00E-02	9.20E-01	8.00E-02	1.97E+11	6.97E-01	3.03E-01
321	NOS3	39495	-99	-99	-99	39803	-99	39231	44	2.32E+05	8.85E-01	1.15E-01	8.85E-01	1.15E-01	1.05E+11	5.56E-01	4.44E-01

Πίνακας 3.1.3: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με τον Διαβήτη Τύπου ΙΙ.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
244	ADRB2	39848	39574	39854	-99	39867	39915	39234	128	981011	7.40E-01	2.60E-01	7.40E-01	2.60E-01	1.22E+13	7.70E-01	2.30E-01
80	IRS2	39684	39456	39924	39867	39914	-99	39532	126	4.15E+05	2.50E-01	7.50E-01	2.50E-01	7.50E-01	7.03E+11	1.60E-01	8.40E-01
61	RETN	39665	-99	39898	-99	39878	39917	39157	116	1.16E+06	9.06E-01	9.40E-02	9.06E-01	9.40E-02	1.12E+13	3.26E-01	6.76E-01
127	CAPN10	39731	-99	39810	-99	39895	39918	39143	116	1.16E+06	9.16E-01	8.40E-02	9.16E-01	8.40E-02	1.12E+13	2.26E-01	7.74E-01
273	BCHE	39877	-99	39915	-99	39864	39912	39674	116	1.16E+06	4.26E-01	5.74E-01	4.26E-01	5.74E-01	1.12E+13	9.92E-01	8.00E-03
86	ABCC8	39690	-99	39821	-99	39921	39916	39139	116	1158734	8.38E-01	1.62E-01	8.38E-01	1.62E-01	1.12E+13	3.06E-01	6.94E-01
85	KCNJ11	39689	-99	39899	-99	39873	39926	39144	116	1.16E+06	6.36E-01	3.66E-01	6.36E-01	3.64E-01	1.12E+13	2.20E-02	9.78E-01
242	ADRB3	39846	-99	39838	39881	39924	-99	39194	114	4.80E+05	2.43E-01	7.57E-01	2.43E-01	7.57E-01	5.78E+11	3.67E-02	9.63E-01
5	GCK	39609	-99	39889	39845	39888	-99	39128	114	4.79E+05	9.10E-01	9.00E-02	9.10E-01	9.00E-02	5.78E+11	1.87E-01	8.13E-01
142	PPARGC1A	39746	-99	39855	39839	39920	-99	39352	114	4.79E+05	5.17E-01	4.83E-01	5.17E-01	4.83E-01	5.78E+11	1.57E-01	8.43E-01
38	CRP	39642	-99	-99	-99	39875	39920	39239	112	1.35E+06	9.60E-01	4.00E-02	9.60E-01	4.00E-02	1.05E+13	1.40E-01	8.60E-01
119	PPARG	39723	-99	-99	-99	39874	39922	39140	112	1346458	9.10E-01	9.00E-02	9.10E-01	9.00E-02	1.05E+13	9.50E-02	9.05E-01
79	IRS1	39683	39259	39812	-99	39905	-99	39131	107	3.60E+05	9.40E-01	6.00E-02	9.40E-01	6.00E-02	6.14E+11	2.00E-02	9.80E-01
82	AKT1	39686	39237	39927	39835	-99	-99	39369	105	2.94E+05	7.70E-01	2.30E-01	7.70E-01	2.30E-01	2.62E+11	1.90E-01	8.10E-01
991	AGER	-99	-99	39820	-99	39871	39921	39447	100	1.40E+06	8.33E-01	1.70E-01	8.33E-01	1.70E-01	1.01E+13	2.90E-01	7.10E-01
986	PAX4	-99	-99	39809	-99	39869	39927	39188	100	1.40E+06	8.43E-01	1.57E-01	8.43E-01	1.57E-01	1.01E+13	3.33E-03	9.97E-01
982	UCP3	-99	-99	39804	-99	39866	39914	39383	100	1.40E+06	9.93E-01	6.67E-03	9.93E-01	6.67E-03	1.01E+13	6.10E-01	3.90E-01
676	STX1A	-99	39625	39837	-99	39870	39919	-99	90	1.40E+06	7.10E-01	2.90E-01	7.10E-01	2.90E-01	1.00E+13	4.60E-01	5.40E-01
102	GCKR	39706	-99	-99	39843	39912	-99	39664	84	4.97E+05	3.80E-01	6.20E-01	3.80E-01	6.20E-01	5.71E+11	3.40E-01	6.60E-01
101	MTHFR	39705	-99	39913	-99	39906	-99	39290	81	4.28E+05	3.76E-01	6.24E-01	3.76E-01	6.24E-01	5.24E+11	1.59E-01	8.41E-01
314	HNF1B	39918	-99	39818	-99	39894	-99	39163	81	4.28E+05	5.67E-01	4.33E-01	5.67E-01	4.33E-01	5.23E+11	4.15E-01	5.85E-01
62	ADIPOQ	39666	-99	39872	-99	39892	-99	39202	81	4.27E+05	8.41E-01	1.59E-01	8.41E-01	1.59E-01	5.24E+11	2.18E-01	7.82E-01
320	IDE	39924	-99	39777	-99	39896	-99	39385	81	4.28E+05	5.19E-01	4.81E-01	5.19E-01	4.81E-01	5.23E+11	5.42E-01	4.58E-01

225	HNF4A	39829	-99	39771	-99	39900	-99	39138	81	4.27E+05	8.76E-01	1.24E-01	8.76E-01	1.24E-01	5.24E+11	2.17E-01	7.83E-01
20	IL6	39624	-99	39906	-99	39918	-99	39189	81	4.28E+05	5.37E-01	4.63E-01	5.37E-01	4.63E-01	5.25E+11	1.60E-02	9.84E-01
315	IGF2BP2	39919	-99	39827	-99	39884	-99	39598	81	4.28E+05	2.66E-01	7.34E-01	2.66E-01	7.34E-01	5.22E+11	8.44E-01	1.56E-01
297	HNF1A	39901	-99	39817	-99	39889	-99	39142	81	4.27E+05	6.77E-01	3.23E-01	6.77E-01	3.23E-01	5.23E+11	4.53E-01	5.47E-01
18	TNF	39622	-99	39904	-99	39913	-99	39185	81	4.28E+05	6.01E-01	3.99E-01	6.01E-01	3.99E-01	5.25E+11	3.50E-02	9.65E-01
70	TCF7L2	39674	-99	39769	-99	39865	-99	39152	81	4.27E+05	1	0	1	0	5.23E+11	5.60E-01	4.40E-01
321	PCSK2	39925	-99	39847	39877	-99	-99	39698	79	3.45E+05	7.00E-02	9.30E-01	7.00E-02	9.30E-01	2.11E+11	7.85E-01	2.15E-01
68	SLC2A2	39672	-99	39881	39854	-99	-99	39232	79	3.44E+05	9.00E-01	1.00E-01	9.00E-01	1.00E-01	2.12E+11	1.95E-01	8.05E-01
83	IGF1R	39687	39391	-99	39840	-99	-99	39384	73	2.65E+05	9.10E-01	9.00E-02	9.10E-01	9.00E-02	2.46E+11	3.90E-01	6.10E-01
81	INSR	39685	39289	39782	-99	-99	-99	39151	70	1.95E+05	9.93E-01	6.67E-03	9.93E-01	6.67E-03	6.88E+10	5.47E-01	4.53E-01
26	TNFRSF1A	39630	39347	39905	-99	-99	-99	39403	70	1.95E+05	7.60E-01	2.40E-01	7.60E-01	2.40E-01	6.92E+10	8.67E-02	9.13E-01
123	PTPN1	39727	39297	39907	-99	-99	-99	39492	70	1.96E+05	5.70E-01	4.30E-01	5.70E-01	4.30E-01	6.92E+10	8.67E-02	9.13E-01
996	CPE	-99	-99	39828	-99	-99	39913	39350	61	1.52E+06	8.60E-01	1.40E-01	8.60E-01	1.40E-01	9.88E+12	4.00E-01	6.00E-01
1048	SLC30A8	-99	-99	39896	-99	39881	-99	39584	60	5.03E+05	3.45E-01	6.55E-01	3.45E-01	6.55E-01	4.54E+11	7.85E-01	2.15E-01
1007	WFS1	-99	-99	39844	-99	39902	-99	39285	60	5.03E+05	6.18E-01	3.83E-01	6.18E-01	3.83E-01	4.56E+11	1.48E-01	8.53E-01
992	CDKAL1	-99	-99	39822	-99	39882	-99	39619	60	5.03E+05	6.38E-01	3.63E-01	6.38E-01	3.63E-01	4.55E+11	7.25E-01	2.75E-01
980	KLF11	-99	-99	39802	-99	39903	-99	39572	60	5.03E+05	5.28E-01	4.73E-01	5.28E-01	4.73E-01	4.55E+11	3.45E-01	6.55E-01
1028	KCNJ9	-99	-99	39869	-99	39868	39925	-99	59	1837495	6.80E-01	3.20E-01	6.80E-01	3.20E-01	7.76E+12	2.30E-01	7.70E-01
1038	INS	-99	-99	39882	39865	-99	-99	39130	58	3.92E+05	7.20E-01	2.80E-01	7.20E-01	2.80E-01	1.85E+11	1.80E-01	8.20E-01
989	GPD2	-99	-99	39815	39885	-99	-99	39496	58	3.92E+05	5.70E-01	4.30E-01	5.70E-01	4.30E-01	1.84E+11	4.45E-01	5.55E-01
599	AGTR1	-99	39543	-99	-99	39927	-99	39135	54	3.98E+05	6.30E-01	3.70E-01	6.30E-01	3.70E-01	5.83E+11	0	1
459	GAPDH	-99	39390	-99	39887	-99	-99	39673	52	2.87E+05	5.40E-01	4.60E-01	5.40E-01	4.60E-01	2.41E+11	5.35E-01	4.65E-01
415	PTPRC	-99	39343	-99	39911	-99	-99	39310	52	2.87E+05	7.45E-01	2.55E-01	7.45E-01	2.55E-01	2.41E+11	4.00E-02	9.60E-01
592	CCR5	-99	39536	39919	-99	-99	-99	39423	49	1.94E+05	2.81E-01	7.19E-01	2.81E-01	7.19E-01	6.91E+10	1.01E-01	8.99E-01
926	MET	-99	39891	39768	-99	-99	-99	39606	49	1.94E+05	4.31E-01	5.69E-01	4.31E-01	5.69E-01	6.82E+10	9.91E-01	8.57E-03
379	SOS1	-99	39299	39789	-99	-99	-99	39766	49	1.93E+05	7.99E-01	2.01E-01	7.99E-01	2.01E-01	6.85E+10	7.90E-01	2.10E-01
797	SELL	-99	39753	39866	-99	-99	-99	39478	49	1.94E+05	2.53E-01	7.47E-01	2.53E-01	7.47E-01	6.88E+10	4.91E-01	5.09E-01

Πίνακας 3.1.4: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με την Υπέρταση.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
19	NPPC	39342	-99	39910	39869	39781	39927	39320	130	6.17E+05	6.30E-01	3.70E-01	6.30E-01	3.70E-01	2.12E+12	1.00E-02	9.90E-01
294	ESR2	39617	39773	39803	-99	39755	39904	39690	128	5.15E+05	6.95E-01	3.05E-01	6.95E-01	3.05E-01	2.31E+12	7.40E-01	2.60E-01
6	EDN1	39329	39822	39856	-99	39765	39905	39133	128	5.15E+05	9.30E-01	7.00E-02	9.30E-01	7.00E-02	2.31E+12	3.75E-01	6.25E-01
4	NPPA	39327	39852	39874	39907	39814	-99	39131	126	3.26E+05	5.50E-01	4.50E-01	5.50E-01	4.50E-01	3.49E+11	5.00E-02	9.50E-01
1190	SCNN1A	-99	-99	39816	39830	39771	39910	39864	123	7.17E+05	8.00E-01	2.00E-01	8.00E-01	2.00E-01	1.82E+12	4.80E-01	5.20E-01
796	AGTR1	-99	39490	39811	-99	39775	39925	39145	121	5.94E+05	8.25E-01	1.75E-01	8.25E-01	1.75E-01	2.12E+12	5.00E-03	9.95E-01
1119	AGT	-99	39862	39794	-99	39746	39890	39132	121	5.94E+05	9.80E-01	2.00E-02	9.80E-01	2.00E-02	2.12E+12	8.10E-01	1.90E-01
7	EDNRA	39330	-99	39824	-99	39750	39892	39191	116	5.97E+05	9.98E-01	1.67E-03	9.98E-01	1.67E-03	2.10E+12	5.13E-01	4.88E-01
459	SLC14A2	39782	-99	39897	-99	39782	39922	39323	116	5.98E+05	2.18E-01	7.82E-01	2.18E-01	7.82E-01	2.11E+12	1.68E-01	8.32E-01
335	PTGIS	39658	-99	39813	-99	39749	39894	39206	116	5.97E+05	9.83E-01	1.67E-02	9.83E-01	1.67E-02	2.10E+12	6.83E-01	3.18E-01
523	RETN	39846	-99	39909	-99	39779	39902	39241	116	5.98E+05	4.62E-01	5.38E-01	4.62E-01	5.38E-01	2.10E+12	6.73E-01	3.27E-01
209	NPR3	39532	-99	39833	-99	39758	39899	39297	116	5.97E+05	9.60E-01	4.00E-02	9.60E-01	4.00E-02	2.10E+12	5.28E-01	4.73E-01
218	HSD11B2	39541	-99	39884	-99	39770	39908	39175	116	5.98E+05	8.14E-01	1.86E-01	8.14E-01	1.86E-01	2.10E+12	2.70E-01	7.30E-01
258	PNMT	39581	-99	39871	-99	39773	39913	39302	116	5.98E+05	6.76E-01	3.24E-01	6.76E-01	3.24E-01	2.11E+12	2.32E-01	7.68E-01
126	APOE	39449	-99	39793	-99	39786	39917	39525	116	5.98E+05	8.41E-01	1.59E-01	8.41E-01	1.59E-01	2.11E+12	1.58E-01	8.42E-01
9	DRD1	39332	-99	39896	-99	39751	39891	39301	116	5.97E+05	9.73E-01	2.67E-02	9.73E-01	2.67E-02	2.10E+12	6.29E-01	3.71E-01
70	ADM	39393	-99	39828	-99	39762	39900	39153	116	5.97E+05	9.93E-01	7.50E-03	9.93E-01	7.50E-03	2.10E+12	3.40E-01	6.60E-01
97	CAT	39420	-99	39814	-99	39757	39896	39533	116	5.97E+05	9.75E-01	2.50E-02	9.75E-01	2.50E-02	2.10E+12	6.52E-01	3.48E-01
182	EDN2	39505	-99	39888	-99	39763	39919	39151	116	5.98E+05	7.27E-01	2.73E-01	7.27E-01	2.73E-01	2.11E+12	3.83E-02	9.62E-01
343	ACE2	39666	-99	39830	39914	39873	-99	39273	114	3.71E+05	5.20E-01	4.80E-01	5.20E-01	4.80E-01	2.89E+11	3.10E-01	6.90E-01
27	NOS3	39350	-99	-99	-99	39774	39921	39154	112	613362	9.05E-01	9.50E-02	9.05E-01	9.50E-02	2.10E+12	1.25E-02	9.88E-01
300	PPARG	39623	-99	-99	-99	39760	39911	39545	112	6.14E+05	6.18E-01	3.83E-01	6.18E-01	3.83E-01	2.10E+12	4.20E-01	5.80E-01
371	PMP22	39694	-99	-99	-99	39756	39893	39139	112	6.13E+05	9.60E-01	4.00E-02	9.60E-01	4.00E-02	2.10E+12	6.98E-01	3.03E-01
33	TNF	39356	-99	-99	-99	39787	39920	39333	112	6.13E+05	7.98E-01	2.03E-01	7.98E-01	2.03E-01	2.10E+12	4.00E-02	9.60E-01
2	BMPR2	39325	39368	39820	-99	39908	-99	39129	107	2.48E+05	9.10E-01	9.00E-02	9.10E-01	9.00E-02	1.67E+11	8.00E-02	9.20E-01
145	CAV1	39468	39253	39807	39890	-99	-99	39363	105	3.11E+05	9.70E-01	3.00E-02	9.70E-01	3.00E-02	3.43E+11	2.20E-01	7.80E-01
1236	SCNN1G	-99	-99	39891	-99	39778	39915	39815	100	7.18E+05	1.80E-01	8.20E-01	1.80E-01	8.20E-01	1.82E+12	5.07E-01	4.93E-01
1200	GNB3	-99	-99	39836	-99	39761	39903	39137	100	7.17E+05	9.82E-01	1.83E-02	9.82E-01	1.83E-02	1.82E+12	2.58E-01	7.42E-01
1229	SCNN1B	-99	-99	39882	-99	39776	39901	39642	100	7.18E+05	5.95E-01	4.05E-01	5.95E-01	4.05E-01	1.82E+12	7.50E-01	2.50E-01

1185	ADD1	-99	-99	39805	-99	39754	39898	39143	100	7.17E+05	9.98E-01	1.67E-03	9.98E-01	1.67E-03	1.82E+12	3.35E-01	6.65E-01
1212	NR3C2	-99	-99	39855	-99	39784	39907	39135	100	7.17E+05	9.02E-01	9.83E-02	9.02E-01	9.83E-02	1.82E+12	2.28E-01	7.72E-01
1253	CYP11B2	-99	-99	39927	-99	39767	39926	39167	100	7.18E+05	3.65E-01	6.35E-01	3.65E-01	6.35E-01	1.82E+12	1.67E-02	9.83E-01
1188	ADD2	-99	-99	39810	39879	39906	-99	39157	98	4.33E+05	6.30E-01	3.70E-01	6.30E-01	3.70E-01	2.09E+11	2.10E-01	7.90E-01
1128	GNAS	-99	39871	-99	-99	39783	39923	39245	96	6.10E+05	3.60E-01	6.40E-01	3.60E-01	6.40E-01	2.12E+12	1.90E-01	8.10E-01
983	ATP1A1	-99	39709	39842	39893	-99	-99	39389	89	3.60E+05	5.20E-01	4.80E-01	5.20E-01	4.80E-01	2.96E+11	5.40E-01	4.60E-01
34	IL6	39357	-99	39912	-99	39799	-99	39309	81	2.84E+05	7.38E-01	2.62E-01	7.38E-01	2.62E-01	1.41E+11	3.11E-02	9.69E-01
382	ACE	39705	-99	39906	-99	39747	-99	39134	81	2.84E+05	7.24E-01	2.76E-01	7.24E-01	2.76E-01	1.41E+11	1.84E-01	8.16E-01
49	CRP	39372	-99	39918	-99	39888	-99	39402	81	2.85E+05	3.30E-01	6.70E-01	3.30E-01	6.70E-01	1.41E+11	4.44E-03	9.96E-01
239	LEP	39562	-99	39899	-99	39877	-99	39136	81	2.84E+05	4.57E-01	5.43E-01	4.57E-01	5.43E-01	1.41E+11	2.44E-02	9.76E-01
270	SELE	39593	-99	39893	-99	39844	-99	39213	81	2.84E+05	5.23E-01	4.77E-01	5.23E-01	4.77E-01	1.41E+11	8.67E-02	9.13E-01
194	ADRB1	39517	-99	39875	-99	39793	-99	39219	81	2.84E+05	8.36E-01	1.64E-01	8.36E-01	1.64E-01	1.41E+11	1.82E-01	8.18E-01
254	COMT	39577	-99	39808	-99	39876	-99	39150	81	2.84E+05	8.50E-01	1.50E-01	8.50E-01	1.50E-01	1.40E+11	4.08E-01	5.92E-01
54	NPR1	39377	-99	39829	-99	39748	-99	39272	81	2.84E+05	9.92E-01	7.78E-03	9.92E-01	7.78E-03	1.40E+11	4.56E-01	5.44E-01
200	MTHFR	39523	-99	39911	-99	39791	-99	39280	81	2.84E+05	6.41E-01	3.59E-01	6.41E-01	3.59E-01	1.41E+11	8.33E-02	9.17E-01
122	TRHR	39445	-99	39908	-99	39785	39918	-99	80	7.25E+05	3.50E-01	6.50E-01	3.50E-01	6.50E-01	1.78E+12	1.70E-01	8.30E-01
329	SLC12A3	39652	-99	39844	39842	-99	-99	39169	79	3.64E+05	8.70E-01	1.30E-01	8.70E-01	1.30E-01	2.87E+11	3.70E-01	6.30E-01
324	CDKN2A	39647	39700	-99	-99	39881	-99	39574	75	1.77E+05	1.32E-01	8.68E-01	1.32E-01	8.68E-01	6.47E+10	2.82E-01	7.18E-01
90	CFTR	39413	39695	-99	-99	39850	-99	39625	75	176580	3.48E-01	6.52E-01	3.48E-01	6.52E-01	6.47E+10	3.54E-01	6.46E-01
271	AR	39594	39300	-99	-99	39872	-99	39681	75	1.77E+05	3.82E-01	6.18E-01	3.82E-01	6.18E-01	6.48E+10	2.08E-01	7.92E-01
100	ESR1	39423	39275	-99	-99	39854	-99	39565	75	1.76E+05	7.50E-01	2.50E-01	7.50E-01	2.50E-01	6.49E+10	1.74E-01	8.26E-01

Πίνακας 3.1.5: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με την Σκλήρυνση κατά πλάκα.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
546	CRYAB	39716	39347	39318	-99	39745	39923	39245	128	6.40E+05	8.90E-01	1.10E-01	8.90E-01	1.10E-01	6.30E+12	5.00E-02	9.50E-01
192	CTLA4	39362	39588	39311	-99	39749	39914	39190	128	6.40E+05	9.85E-01	1.50E-02	9.85E-01	1.50E-02	6.30E+12	3.10E-01	6.90E-01
3	MBP	39173	39505	39303	39926	39884	-99	39142	126	3.80E+06	9.00E-01	1.00E-01	9.00E-01	1.00E-01	3.92E+14	3.00E-02	9.70E-01
1324	CD24	-99	-99	39301	39924	39753	39927	39351	123	5094257	9.00E-01	1.00E-01	9.00E-01	1.00E-01	3.57E+14	3.10E-01	6.90E-01
779	PRKCA	-99	39258	39298	-99	39733	39916	39349	121	7.45E+05	9.90E-01	1.00E-02	9.90E-01	1.00E-02	5.97E+12	1.30E-01	8.70E-01
90	VDR	39260	-99	39527	-99	39739	39912	39165	116	7.43E+05	9.85E-01	1.50E-02	9.85E-01	1.50E-02	5.98E+12	2.95E-01	7.05E-01
7	IL6	39177	-99	39296	-99	39761	39913	39164	116	742968	1	0	1	0	5.98E+12	1.25E-01	8.75E-01
20	IL4	39190	-99	39445	-99	39751	39922	39169	116	7.43E+05	9.65E-01	3.50E-02	9.65E-01	3.50E-02	5.98E+12	2.25E-02	9.78E-01
18	APOE	39188	-99	39289	-99	39760	39921	39175	116	7.43E+05	9.93E-01	7.50E-03	9.93E-01	7.50E-03	5.98E+12	1.50E-02	9.85E-01
105	IL4R	39275	-99	-99	-99	39757	39919	39292	112	8.95E+05	8.50E-01	1.50E-01	8.50E-01	1.50E-01	5.52E+12	6.00E-02	9.40E-01
266	CD80	39436	39464	39406	-99	39858	-99	39259	107	1.85E+05	8.10E-01	1.90E-01	8.10E-01	1.90E-01	8.52E+10	1.42E-01	8.58E-01
35	CD28	39205	39356	39390	-99	39838	-99	39280	107	1.85E+05	9.57E-01	4.33E-02	9.57E-01	4.33E-02	8.52E+10	1.32E-01	8.68E-01
8	IL1A	39178	39533	39501	-99	39898	-99	39677	107	1.85E+05	4.48E-01	5.52E-01	4.48E-01	5.52E-01	8.53E+10	9.78E-02	9.02E-01
267	CD86	39437	39455	39345	-99	39839	-99	39254	107	1.85E+05	8.83E-01	1.17E-01	8.83E-01	1.17E-01	8.51E+10	2.06E-01	7.94E-01
81	IL1RN	39251	39800	39316	-99	39840	-99	39218	107	1.85E+05	8.34E-01	1.66E-01	8.34E-01	1.66E-01	8.51E+10	2.26E-01	7.74E-01
42	IL12B	39212	39577	39343	-99	39914	-99	39160	107	1.85E+05	8.43E-01	1.57E-01	8.43E-01	1.57E-01	8.56E+10	4.44E-03	9.96E-01
9	IL1B	39179	39494	39315	-99	39823	-99	39189	107	1.85E+05	9.72E-01	2.78E-02	9.72E-01	2.78E-02	8.51E+10	1.71E-01	8.29E-01
279	MMP9	39449	39874	39459	-99	39865	-99	39174	107	1.85E+05	4.57E-01	5.43E-01	4.57E-01	5.43E-01	8.51E+10	2.22E-01	7.78E-01
129	BDNF	39299	39707	39349	-99	39885	-99	39140	107	1.85E+05	7.74E-01	2.26E-01	7.74E-01	2.26E-01	8.54E+10	5.22E-02	9.48E-01
1424	WT1	-99	-99	39495	-99	39748	39917	39266	100	9.01E+05	8.23E-01	1.77E-01	8.23E-01	1.77E-01	5.48E+12	1.77E-01	8.27E-01
1331	SH2D2A	-99	-99	39319	-99	39759	39924	39463	100	9.01E+05	7.10E-01	2.90E-01	7.10E-01	2.90E-01	5.48E+12	1.33E-02	9.87E-01
1748	COL1A1	-99	-99	39902	-99	39743	39898	39751	100	9.01E+05	5.57E-01	4.43E-01	5.57E-01	4.43E-01	5.47E+12	9.83E-01	1.67E-02
848	CREBBP	-99	39351	-99	-99	39741	39900	39210	96	8.97E+05	9.90E-01	1.00E-02	9.90E-01	1.00E-02	5.50E+12	4.00E-01	6.00E-01
1123	CNTF	-99	39695	39310	-99	39915	-99	39214	91	2.03E+05	5.57E-01	4.43E-01	5.57E-01	4.43E-01	7.91E+10	3.00E-02	9.70E-01
787	TP53	-99	39269	39889	-99	39848	-99	39306	91	2.03E+05	5.00E-01	5.00E-01	5.00E-01	5.00E-01	7.87E+10	2.53E-01	7.47E-01
1210	CCR5	-99	39804	39290	-99	39787	-99	39192	91	2.03E+05	8.27E-01	1.73E-01	8.27E-01	1.73E-01	7.84E+10	4.80E-01	5.20E-01
91	CIITA	39261	-99	-99	39925	39861	-99	39159	84	5633847	8.60E-01	1.40E-01	8.60E-01	1.40E-01	3.52E+14	1.10E-01	8.90E-01
144	FAS	39314	-99	39299	-99	39874	-99	39675	81	2.01E+05	5.99E-01	4.01E-01	5.99E-01	4.01E-01	8.04E+10	1.49E-01	8.51E-01
521	IFNAR2	39691	-99	39651	-99	39788	-99	39170	81	2.01E+05	6.16E-01	3.84E-01	6.16E-01	3.84E-01	7.99E+10	6.06E-01	3.94E-01

66	MTHFR	39236	-99	39484	-99	39769	-99	39705	81	2.01E+05	7.86E-01	2.14E-01	7.86E-01	2.14E-01	7.98E+10	6.72E-01	3.28E-01
336	FASLG	39506	-99	39363	-99	39917	-99	39270	81	2.01E+05	5.59E-01	4.41E-01	5.59E-01	4.41E-01	8.07E+10	1.96E-02	9.80E-01
14	LTA	39184	-99	39375	-99	39816	-99	39231	81	2.00E+05	9.69E-01	3.13E-02	9.69E-01	3.13E-02	8.04E+10	1.96E-01	8.04E-01
22	IL2	39192	-99	39475	-99	39860	-99	39168	81	2.00E+05	8.99E-01	1.01E-01	8.99E-01	1.01E-01	8.06E+10	5.65E-02	9.43E-01
503	SOD2	39673	-99	39921	-99	39785	-99	39694	81	2.01E+05	7.57E-02	9.24E-01	7.57E-02	9.24E-01	7.95E+10	9.24E-01	7.61E-02
12	TNFRSF1A	39182	-99	39381	-99	39778	-99	39257	81	2.00E+05	9.87E-01	1.35E-02	9.87E-01	1.35E-02	8.01E+10	3.66E-01	6.34E-01
257	TAC1	39427	-99	39336	-99	39836	-99	39560	81	2.01E+05	6.88E-01	3.12E-01	6.88E-01	3.12E-01	8.02E+10	3.17E-01	6.83E-01
252	CNR1	39422	-99	39593	-99	39924	-99	39167	81	2.01E+05	4.60E-01	5.40E-01	4.60E-01	5.40E-01	8.08E+10	1.57E-02	9.84E-01
65	ICAM1	39235	-99	39355	-99	39821	-99	39187	81	2.00E+05	9.67E-01	3.26E-02	9.67E-01	3.26E-02	8.04E+10	1.71E-01	8.29E-01
11	IFNG	39181	-99	39293	-99	39869	-99	39156	81	2.00E+05	9.60E-01	4.04E-02	9.60E-01	4.04E-02	8.07E+10	2.09E-02	9.79E-01
301	TGFB1	39471	-99	39425	-99	39856	-99	39671	81	2.01E+05	4.05E-01	5.95E-01	4.05E-01	5.95E-01	8.02E+10	3.24E-01	6.76E-01
41	MOG	39211	-99	39291	-99	39897	-99	39162	81	2.00E+05	9.20E-01	8.04E-02	9.20E-01	8.04E-02	8.09E+10	2.17E-03	9.98E-01
10	IL10	39180	-99	39295	-99	39809	-99	39158	81	2.00E+05	9.95E-01	5.22E-03	9.95E-01	5.22E-03	8.04E+10	1.67E-01	8.33E-01
57	CCL2	39227	-99	39392	-99	39844	-99	39178	81	2.00E+05	9.36E-01	6.39E-02	9.36E-01	6.39E-02	8.05E+10	8.87E-02	9.11E-01
245	NOS3	39415	-99	39489	-99	39841	-99	39328	81	2.01E+05	7.07E-01	2.93E-01	7.07E-01	2.93E-01	8.03E+10	2.56E-01	7.44E-01
481	SPP1	39651	-99	39314	-99	39862	-99	39230	81	2.01E+05	6.95E-01	3.05E-01	6.95E-01	3.05E-01	8.04E+10	1.62E-01	8.38E-01
6	TNF	39176	-99	39292	-99	39780	-99	39155	81	2.00E+05	1	0	1	0	8.02E+10	2.82E-01	7.18E-01
108	CD40	39278	-99	39447	-99	39791	-99	39258	81	2.00E+05	9.37E-01	6.26E-02	9.37E-01	6.26E-02	8.01E+10	3.69E-01	6.31E-01
706	TLR4	39876	-99	39348	-99	39770	-99	39721	81	2.01E+05	3.83E-01	6.17E-01	3.83E-01	6.17E-01	7.95E+10	8.91E-01	1.09E-01
658	NFKB1	39828	-99	39397	-99	39794	-99	39744	81	2.01E+05	2.86E-01	7.14E-01	2.86E-01	7.14E-01	7.96E+10	8.11E-01	1.89E-01
203	PTGS2	39373	-99	39502	-99	39764	-99	39700	81	2.01E+05	6.90E-01	3.10E-01	6.90E-01	3.10E-01	7.97E+10	7.69E-01	2.31E-01
308	NR3C1	39478	39904	-99	-99	39916	-99	39734	75	1.99E+05	2.25E-02	9.78E-01	2.25E-02	9.78E-01	8.19E+10	2.70E-01	7.30E-01

Πίνακας 3.1.6: Αποτελέσματα Metradisc με τα πρώτα 50 υψηλά υποψήφια γονίδια (top genes) που σχετίζονται με την Παχυσαρκία.

gene id	gene	S1	S2	S3	S4	S5	S6	S7	Κλάση	R-mean	right-sided P-value for R-mean	left-sided P-value for R-mean	right-sided P-value for R-weighted mean	left-sided P-value for R-weighted mean	Q-mean	right-sided P-value for Q-mean	left-sided P-value for Q-mean
39	MC4R	39409	39714	39573	-99	39771	39927	39166	128	727614	8.93E-01	1.08E-01	8.93E-01	1.08E-01	8.78E+12	2.50E-03	9.98E-01
98	PPARG	39468	39549	39576	-99	39773	39921	39131	128	7.27E+05	9.80E-01	2.00E-02	9.80E-01	2.00E-02	8.78E+12	6.75E-02	9.33E-01
85	NR3C1	39455	39369	39698	-99	39777	39912	39225	128	7.27E+05	9.95E-01	5.00E-03	9.95E-01	5.00E-03	8.77E+12	5.30E-01	4.70E-01
139	ADRB2	39509	39439	39617	-99	39762	39913	39140	128	7.27E+05	9.98E-01	2.50E-03	9.98E-01	2.50E-03	8.77E+12	4.38E-01	5.63E-01
596	NR0B2	-99	39297	39611	-99	39783	39923	39262	121	8.48E+05	9.73E-01	2.67E-02	9.73E-01	2.67E-02	8.35E+12	6.67E-03	9.93E-01
945	SORBS1	-99	39702	39587	-99	39782	39926	39383	121	8.48E+05	7.03E-01	2.97E-01	7.03E-01	2.97E-01	8.35E+12	3.67E-02	9.63E-01
1094	ADRA2A	-99	39864	39707	-99	39780	39918	39354	121	8.48E+05	6.20E-01	3.80E-01	6.20E-01	3.80E-01	8.34E+12	5.23E-01	4.80E-01
28	POMC	39398	-99	39634	-99	39772	39914	39134	116	8.53E+05	9.98E-01	2.22E-03	9.98E-01	2.22E-03	8.30E+12	2.97E-01	7.03E-01
35	AGRP	39405	-99	39627	-99	39776	39920	39144	116	8.53E+05	9.83E-01	1.67E-02	9.83E-01	1.67E-02	8.31E+12	4.89E-02	9.51E-01
155	LIPE	39525	-99	39596	-99	39785	39922	39143	116	8.53E+05	9.57E-01	4.33E-02	9.57E-01	4.33E-02	8.31E+12	3.67E-02	9.63E-01
11	TNF	39381	-99	39575	-99	39786	39924	39172	116	8.53E+05	9.77E-01	2.33E-02	9.77E-01	2.33E-02	8.31E+12	4.44E-03	9.96E-01
53	UCP2	39423	-99	39590	-99	39761	39910	39136	116	8.52E+05	1	0	1	0	8.30E+12	4.63E-01	5.37E-01
232	MC3R	39602	-99	39637	-99	39784	39925	39130	116	8.53E+05	8.53E-01	1.47E-01	8.53E-01	1.47E-01	8.31E+12	1.22E-02	9.88E-01
5	LPL	39375	-99	39660	-99	39765	39908	39138	116	8.52E+05	1	0	1	0	8.30E+12	5.92E-01	4.08E-01
4	LEP	39374	-99	39572	-99	39778	39919	39128	116	8.53E+05	9.97E-01	3.33E-03	9.97E-01	3.33E-03	8.31E+12	4.67E-02	9.53E-01
52	UCP3	39422	-99	39594	-99	39766	39911	39186	116	8.52E+05	9.99E-01	1.11E-03	9.99E-01	1.11E-03	8.30E+12	4.39E-01	5.61E-01
14	IL1B	39384	-99	39724	39868	39861	-99	39557	114	3.11E+05	5.75E-01	4.25E-01	5.75E-01	4.25E-01	2.69E+11	1.45E-01	8.55E-01
33	NPY	39403	-99	39629	39831	39881	-99	39183	114	3.11E+05	9.55E-01	4.50E-02	9.55E-01	4.50E-02	2.69E+11	1.60E-01	8.40E-01
114	HTR2A	39484	39743	-99	39859	39888	-99	39294	110	2.92E+05	4.40E-01	5.60E-01	4.40E-01	5.60E-01	2.98E+11	2.90E-01	7.10E-01
221	PPARGC1A	39591	39337	39672	-99	39839	-99	39149	107	1.89E+05	9.26E-01	7.44E-02	9.26E-01	7.44E-02	7.95E+10	2.28E-01	7.72E-01
109	ESR1	39479	39241	39699	-99	39793	-99	39556	107	1.89E+05	9.33E-01	6.78E-02	9.33E-01	6.67E-02	7.91E+10	5.67E-01	4.33E-01
166	VDR	39536	39457	39763	-99	39833	-99	39697	107	1.90E+05	4.67E-01	5.33E-01	4.67E-01	5.33E-01	7.92E+10	5.37E-01	4.63E-01
99	PPARD	39469	39795	39582	-99	39837	-99	39151	107	1.89E+05	8.81E-01	1.19E-01	8.81E-01	1.19E-01	7.93E+10	3.99E-01	6.01E-01
177	IRS1	39547	39264	39663	-99	39809	-99	39220	107	1.89E+05	9.71E-01	2.89E-02	9.71E-01	2.89E-02	7.93E+10	3.61E-01	6.39E-01
119	PTPN1	39489	39296	39795	-99	39887	-99	39248	107	1.89E+05	7.49E-01	2.51E-01	7.49E-01	2.51E-01	7.97E+10	4.11E-02	9.59E-01
178	IRS2	39548	39534	39584	-99	39769	-99	39258	107	1.89E+05	9.77E-01	2.33E-02	9.77E-01	2.33E-02	7.90E+10	7.44E-01	2.56E-01
160	AR	39530	39265	39743	-99	39864	-99	39314	107	1.89E+05	8.32E-01	1.68E-01	8.32E-01	1.68E-01	7.96E+10	1.27E-01	8.73E-01
97	PPARA	39467	39548	39684	-99	39859	-99	39212	107	1.89E+05	8.42E-01	1.58E-01	8.42E-01	1.58E-01	7.95E+10	2.02E-01	7.98E-01
179	INSR	39549	39325	39671	39881	-99	-99	39184	105	2.45E+05	9.15E-01	8.50E-02	9.15E-01	8.50E-02	2.78E+11	4.00E-02	9.60E-01

266	NR1H3	39636	39640	39602	39856	-99	-99	39658	105	2.46E+05	5.95E-01	4.05E-01	5.95E-01	4.05E-01	2.77E+11	8.00E-01	2.00E-01
1184	SIM1	-99	-99	39639	-99	39767	39915	39231	100	1.03E+06	9.53E-01	4.67E-02	9.53E-01	4.67E-02	7.65E+12	1.60E-01	8.40E-01
1186	PCSK1	-99	-99	39641	-99	39764	39907	39142	100	1.03E+06	9.93E-01	6.67E-03	9.93E-01	6.67E-03	7.64E+12	5.03E-01	4.97E-01
1193	GNB3	-99	-99	39651	-99	39774	39916	39146	100	1.03E+06	9.47E-01	5.33E-02	9.47E-01	5.33E-02	7.65E+12	1.30E-01	8.70E-01
1187	CIDEA	-99	-99	39643	39861	39916	-99	39157	98	3.57E+05	7.40E-01	2.60E-01	7.40E-01	2.60E-01	2.27E+11	3.50E-02	9.65E-01
1154	LEPR	-99	-99	39579	39864	39775	-99	39132	98	3.56E+05	9.95E-01	5.00E-03	9.95E-01	5.00E-03	2.27E+11	4.50E-02	9.55E-01
668	NFKB1	-99	39387	39874	-99	39895	-99	39648	91	2.05E+05	1.33E-01	8.68E-01	1.33E-01	8.68E-01	7.47E+10	2.00E-01	8.00E-01
865	CEBPA	-99	39612	39701	-99	39914	-99	39463	91	2.05E+05	2.40E-01	7.60E-01	2.40E-01	7.60E-01	7.48E+10	1.33E-01	8.68E-01
1142	CNR1	-99	39915	39916	-99	39868	-99	39208	91	2.05E+05	8.50E-02	9.15E-01	8.50E-02	9.15E-01	7.45E+10	3.70E-01	6.30E-01
832	AGTR1	-99	39579	39847	-99	39897	-99	39614	91	2.05E+05	8.25E-02	9.18E-01	8.25E-02	9.18E-01	7.47E+10	2.65E-01	7.35E-01
140	HSD11B1	39510	-99	39624	-99	39829	-99	39139	81	2.11E+05	9.67E-01	3.32E-02	9.67E-01	3.32E-02	6.98E+10	2.51E-01	7.49E-01
37	RETN	39407	-99	39642	-99	39781	-99	39133	81	2.11E+05	9.98E-01	2.11E-03	9.98E-01	2.11E-03	6.96E+10	4.47E-01	5.53E-01
74	CYP19A1	39444	-99	39657	-99	39830	-99	39546	81	2.11E+05	8.18E-01	1.82E-01	8.18E-01	1.82E-01	6.96E+10	4.65E-01	5.35E-01
43	ADRB3	39413	-99	39599	-99	39768	-99	39169	81	2.11E+05	1	0	1	0	6.95E+10	5.37E-01	4.63E-01
173	HTR2C	39543	-99	39670	-99	39827	-99	39293	81	2.11E+05	8.70E-01	1.30E-01	8.70E-01	1.30E-01	6.96E+10	3.84E-01	6.16E-01
472	AGT	39842	-99	39678	-99	39808	-99	39286	81	2.11E+05	6.42E-01	3.58E-01	6.42E-01	3.58E-01	6.94E+10	6.77E-01	3.23E-01
12	IL6	39382	-99	39621	-99	39794	-99	39175	81	2.11E+05	9.97E-01	2.63E-03	9.97E-01	2.63E-03	6.96E+10	3.81E-01	6.19E-01
205	CAPN10	39575	-99	39597	-99	39826	-99	39334	81	2.11E+05	9.06E-01	9.42E-02	9.06E-01	9.42E-02	6.96E+10	4.29E-01	5.71E-01
7	APOB	39377	-99	39668	-99	39865	-99	39187	81	2.11E+05	9.39E-01	6.08E-02	9.39E-01	6.08E-02	7.00E+10	7.08E-02	9.29E-01
377	NOS3	39747	-99	39721	-99	39835	-99	39574	81	2.12E+05	3.44E-01	6.56E-01	3.44E-01	6.56E-01	6.94E+10	6.60E-01	3.40E-01
465	INSIG2	39835	-99	39601	-99	39792	-99	39587	81	2.11E+05	6.12E-01	3.88E-01	6.12E-01	3.88E-01	6.92E+10	9.04E-01	9.55E-02

3.2 Αποτελέσματα R-PROJECT

Στις επόμενες σελίδες που ακολουθούν παρουσιάζονται τα αποτελέσματα που είχαμε εξάγει κάνοντας χρήση το πακέτο στατιστικής ανάλυσης RankProd του προγράμματος R-Project. Τα αποτελέσματα αναλύονται μέσα από τις τρείς βασικές συναρτήσεις που μας προσφέρει το πακέτο RankProd που είναι:

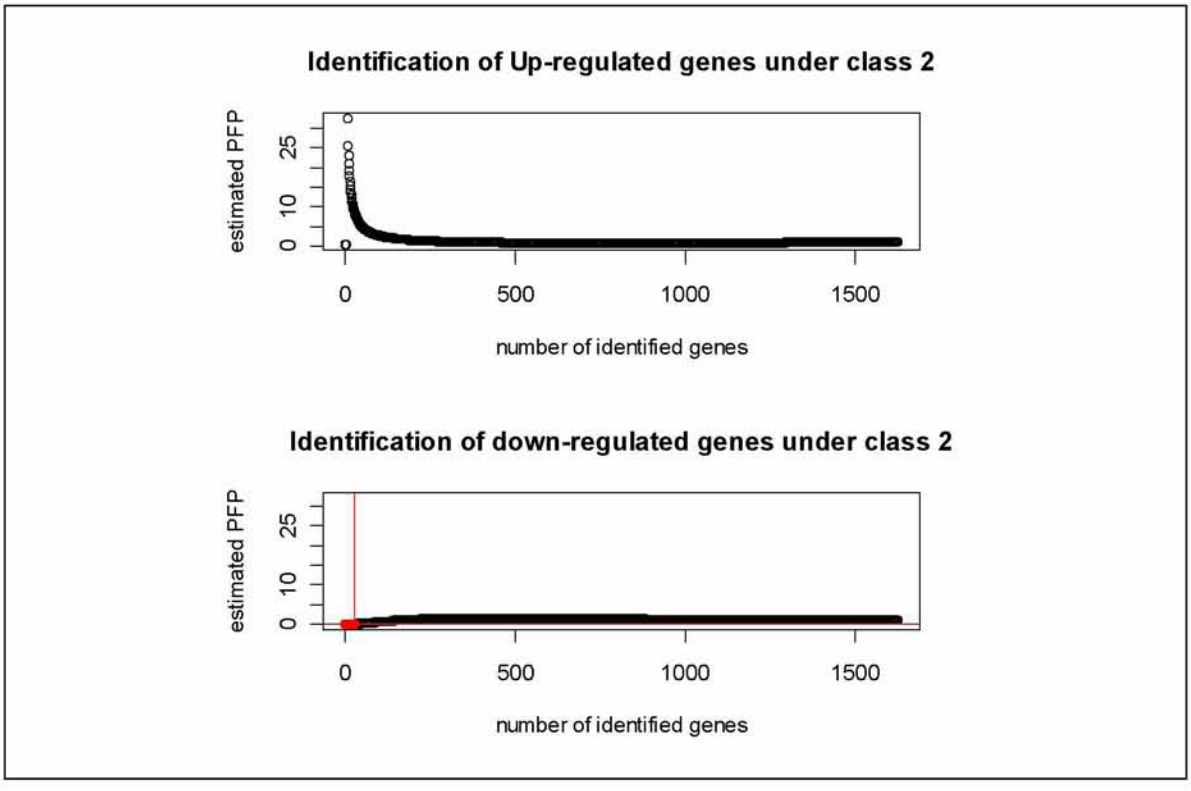
-Παρουσίαση των αποτελεσμάτων σε γραφική παράσταση με βάση την συνάρτηση «plotRP» με δεδομένα εισόδου RP.out και ποσοστό ψευδώς θετικών προβλέψεων «pfp» να είναι μικρότερο του 0.05. (cutoff = 0.05)

-Παρουσίαση των αποτελεσμάτων σε γραφική παράσταση με βάση την συνάρτηση «plotRP» με δεδομένα εισόδου RP.out=RPadvance και ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05. (cutoff = 0.05)

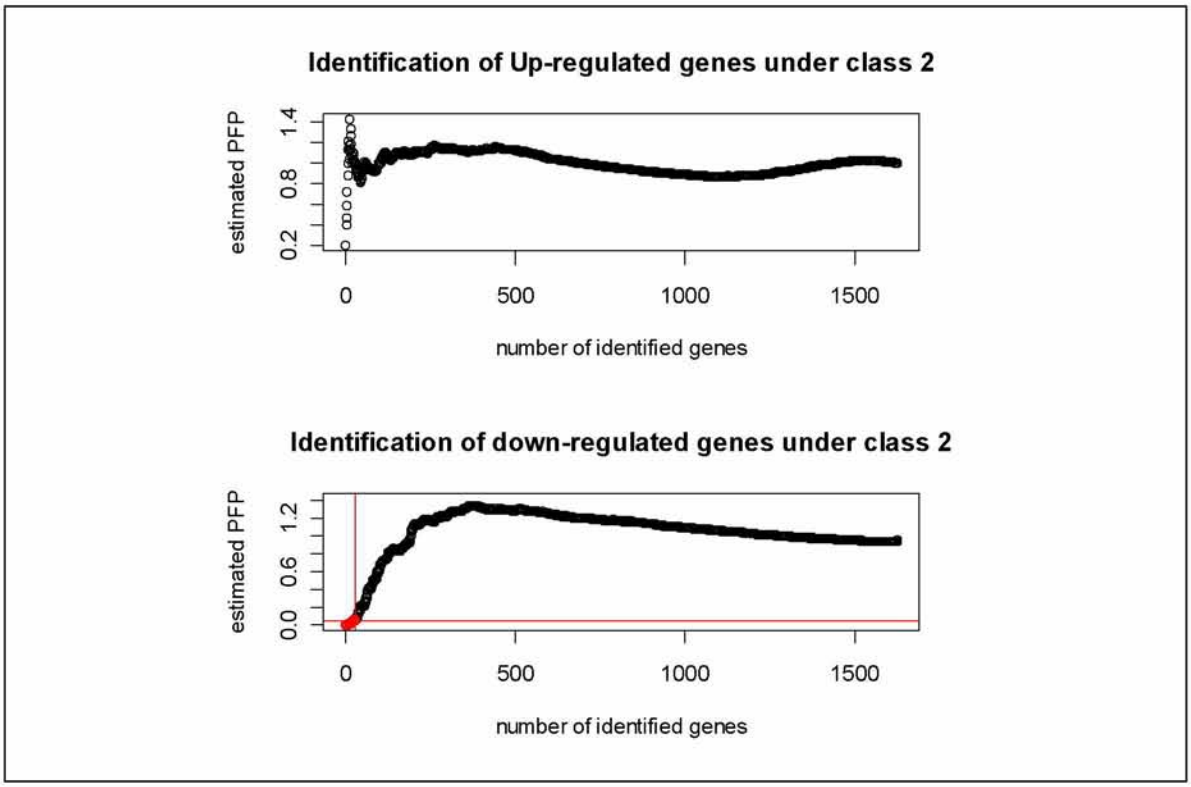
-Παρουσίαση των υψηλά υποψήφιων γονιδίων (top genes)που σχετίζονται με μία ασθένεια με την συνάρτηση «topGene» με ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05. (cutoff = 0.05)

3.2.1 Αποτελέσματα για Καρκίνο μαστού.

Εικόνα 3.2.1.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Καρκίνο μαστού.



Εικόνα 3.2.1.2: Αποτελέσματα συνάρτησης plotRP με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05 για Καρκίνο μαστού όπου RP.out=RPadvance. (RP.out, cutoff = 0.05)



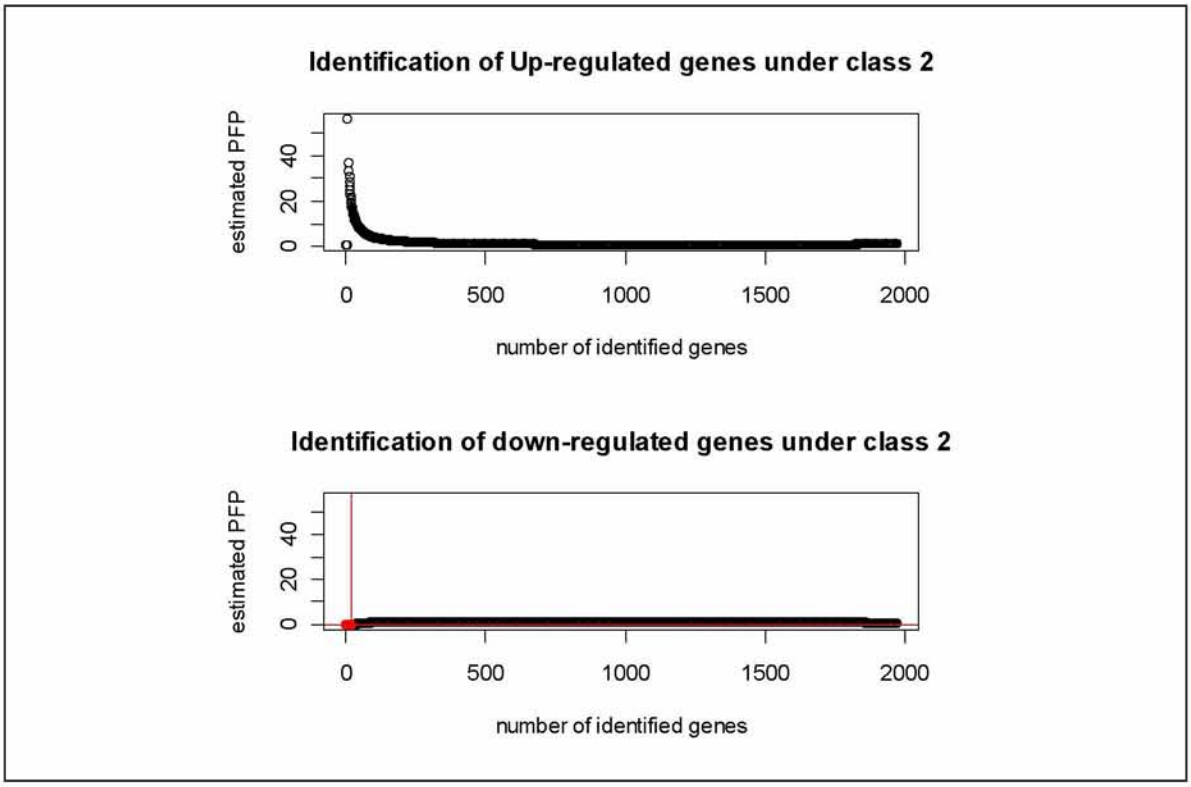
Εικόνα 3.2.1.3: Αποτελέσματα συνάρτησης topGene() για Καρκίνο μαστού με cutoff = 0.05.

```
Table2: Genes called significant under class1 > class2

$Table1
NULL
$Table2
        gene.index  RP/Rsum FC:(class1/class2)    pfp
BRCA1            8  10.7282                 NA 0.0000
BRCA2            6  29.4108                 NA 0.0000
ERBB2           24  31.5644                 NA 0.0000
ESR1            25  35.4800                 NA 0.0000
TP53           387  35.5745                 NA 0.0000
ATM            277 108.6115                 NA 0.0017
SRC            251 135.0436                 NA 0.0129
CDH1            70 136.3081                 NA 0.0112
CYP19A1         39 140.2569                 NA 0.0111
VEGFA           31 148.8821                 NA 0.0150
VDR             86 151.7693                 NA 0.0145
EGFR            29 153.4884                 NA 0.0133
PTEN           564 153.6701                 NA 0.0123
MYC             79 156.3512                 NA 0.0121
RAD51           64 156.5159                 NA 0.0120
NCOA3          202 164.6249                 NA 0.0175
XRCC3          604 171.3363                 NA 0.0271
PIK3R1          58 172.3127                 NA 0.0267
```

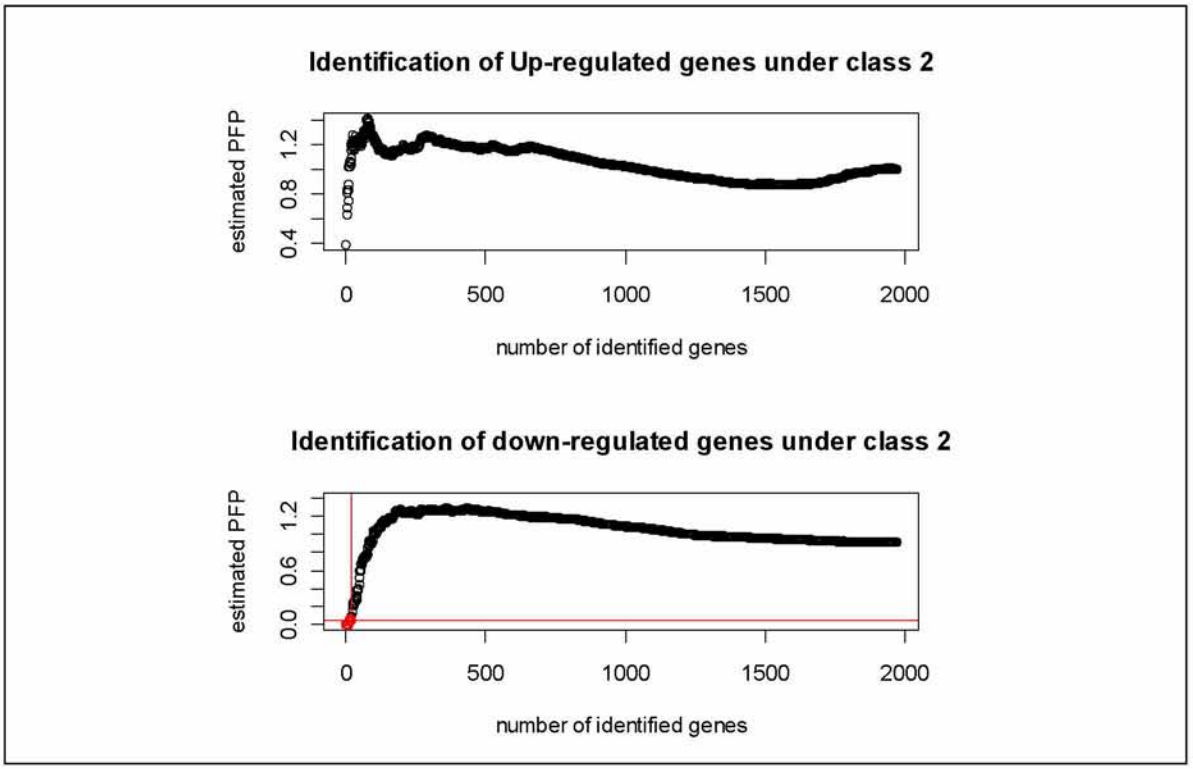
AR	231	173.5059	NA 0.0263
CHEK2	334	179.1722	NA 0.0280
BARD1	45	181.4831	NA 0.0281
CCND1	108	195.4255	NA 0.0459
MDM2	138	202.6300	NA 0.0565
SNCG	21	203.1116	NA 0.0554
CDKN1A	137	204.2743	NA 0.0552
SHBG	40	204.6497	NA 0.0542
STAT3	154	204.9258	NA 0.0526
CYP1A1	358	205.1102	NA 0.0511
ABCG2	591	205.9502	NA 0.0497

3.2.2 Αποτελέσματα για Διαβήτη τύπου Ι.

Εικόνα 3.2.2.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι.



Εικόνα 3.2.2.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι όπου RP.out=RPadvance .

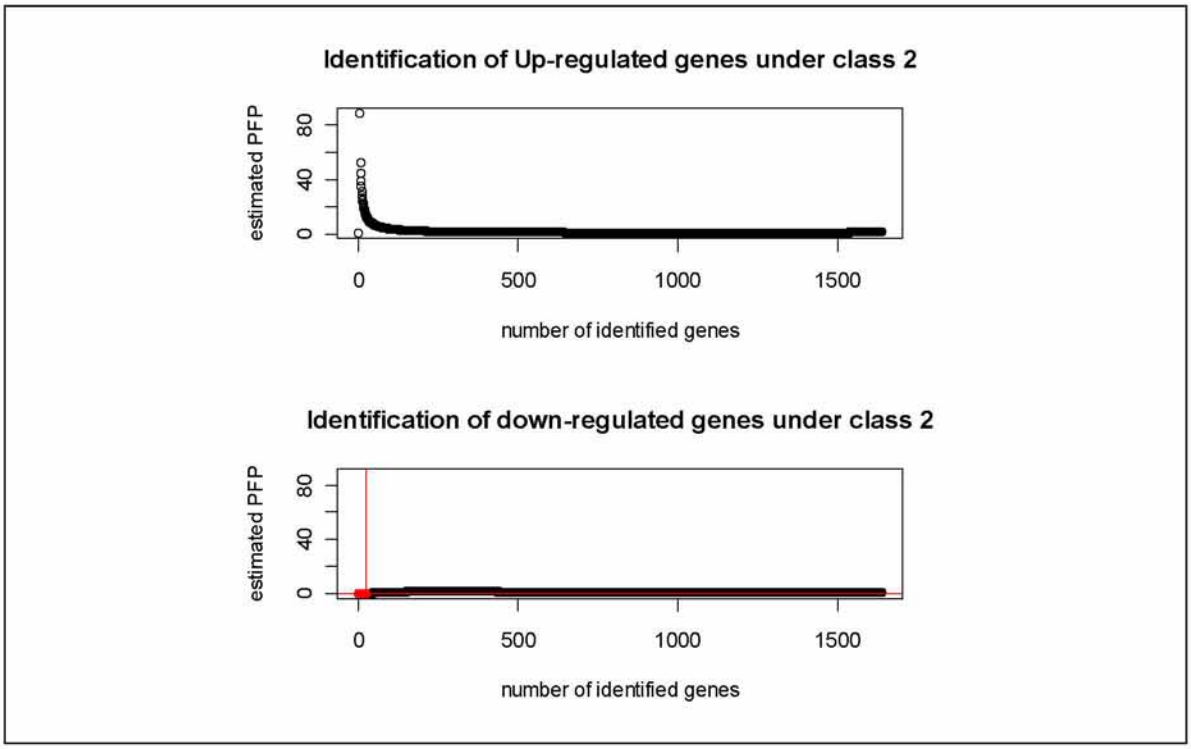


Εικόνα 3.2.2.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη Ι με cutoff = 0.05 .

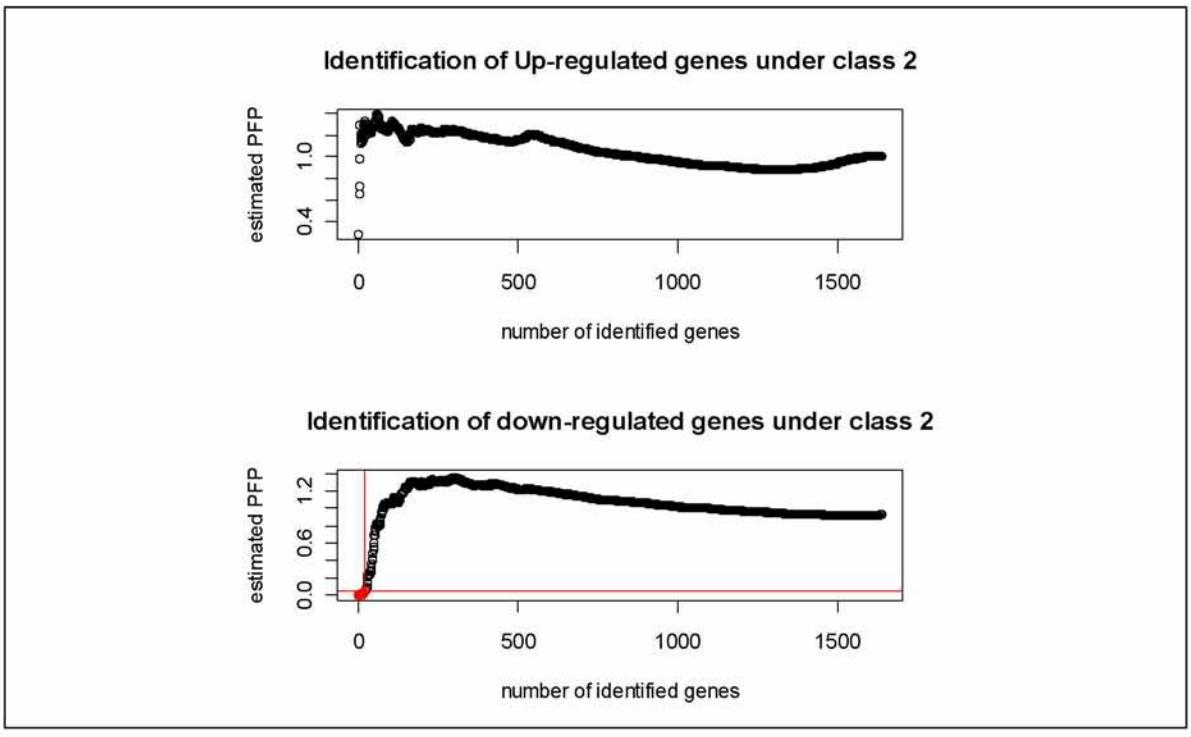
```
Table2: Genes called significant under class1 > class2
$Table1
NULL
$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
CTLA4          63  52.9042                 NA 0.0000   0e+00
VDR            77  64.2509                 NA 0.0000   0e+00
IL6            26  73.9151                 NA 0.0000   0e+00
IRS1           44  88.7848                 NA 0.0000   0e+00
PTPN22        181  97.0429                 NA 0.0000   0e+00
SUMO4         779 113.5819                 NA 0.0000   0e+00
IRS2           45 128.3739                 NA 0.0000   0e+00
TAP2          588 128.5875                 NA 0.0000   0e+00
ICAM1          40 159.2928                 NA 0.0044   0e+00
TNF            24 165.9355                 NA 0.0050   0e+00
WFS1          597 199.3906                 NA 0.0255   1e-04
PPARG         185 201.4936                 NA 0.0258   2e-04
VEGFA         162 206.2068                 NA 0.0269   2e-04
HNF1A         227 216.6946                 NA 0.0379   3e-04
NEUROD1       158 231.0030                 NA 0.0547   4e-04
IL18           93 231.3577                 NA 0.0512   4e-04
RETN           82 231.9293                 NA 0.0506   4e-04
GCK           240 233.4505                 NA 0.0494   5e-04
```

3.2.3 Αποτελέσματα για Διαβήτη τύπου ΙΙ.

Εικόνα 3.2.3.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ.



Εικόνα 3.2.3.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ όπου RP.out=RPadvance.

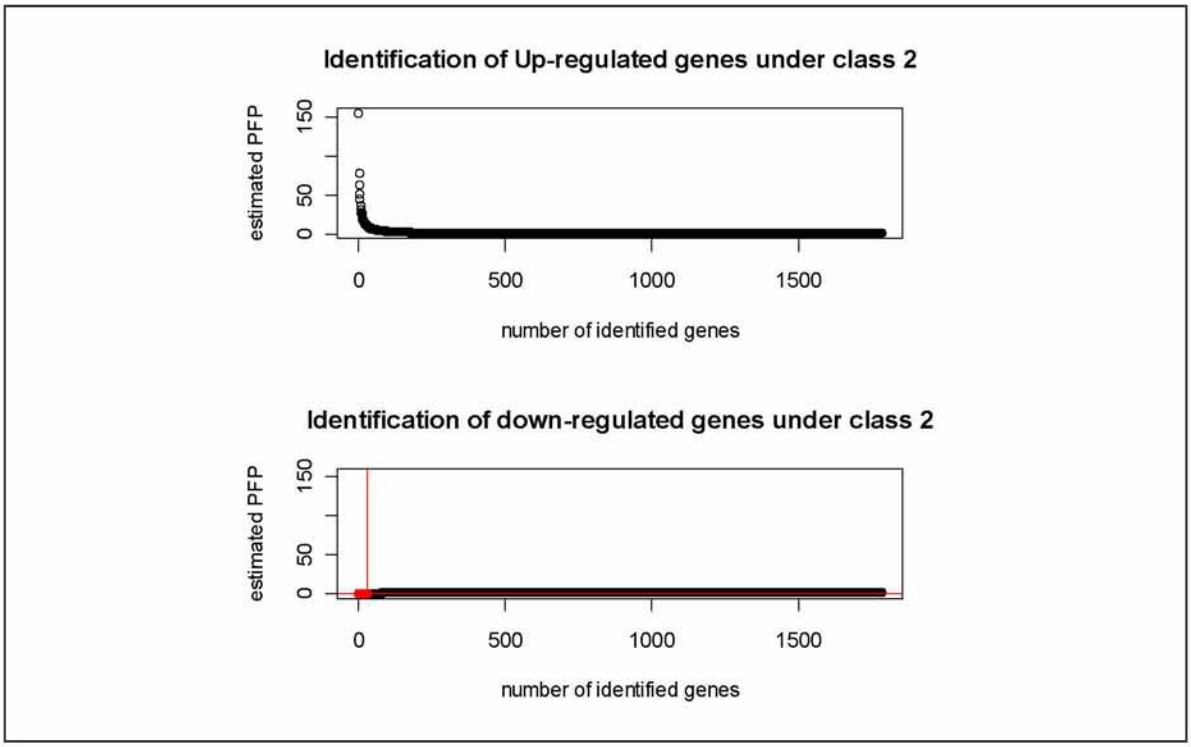


Εικόνα 3.2.3.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη II με cutoff = 0.05 .

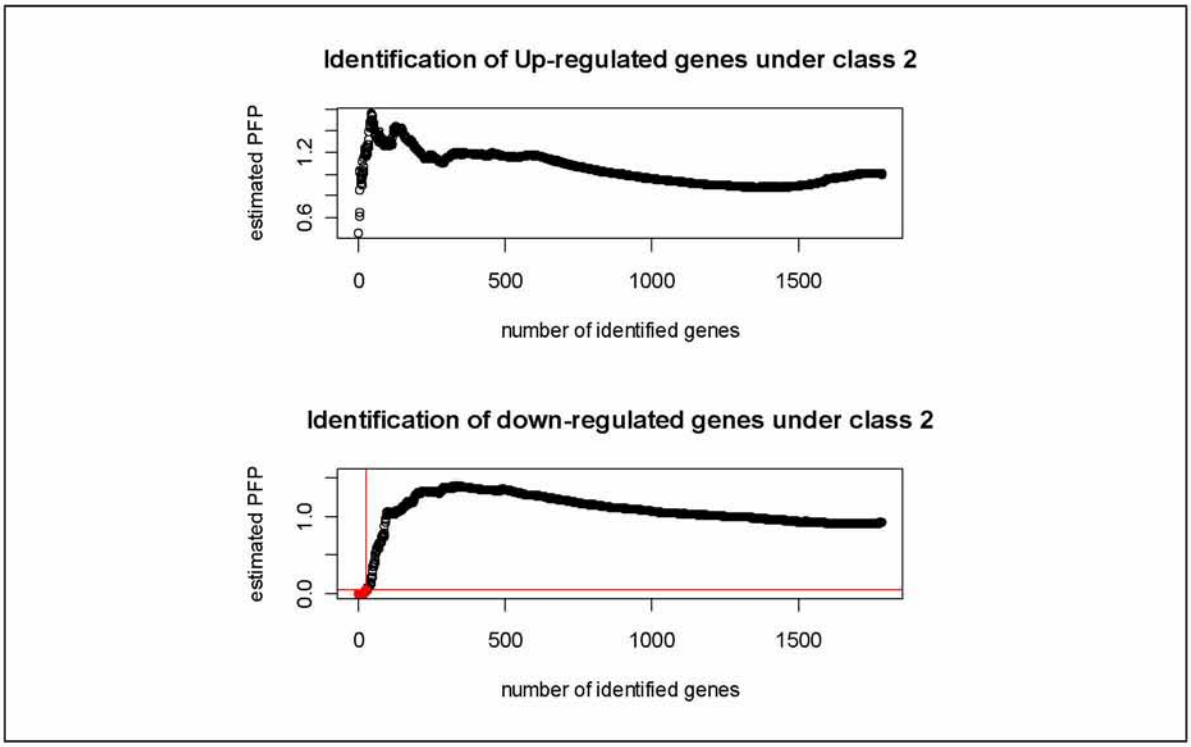
```
Table2: Genes called significant under class1 > class2
$Table1
NULL
$Table2
         gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
GCK               5  95.3130                 NA 0.0000   0e+00
PPARG           119 101.7111                 NA 0.0000   0e+00
KCNJ11           85 102.4308                 NA 0.0000   0e+00
RETN             61 111.0486                 NA 0.0000   0e+00
TCF7L2           70 112.1429                 NA 0.0020   0e+00
PPARGC1A        142 128.7652                 NA 0.0050   0e+00
UCP3            352 137.8269                 NA 0.0057   0e+00
BCHE            273 141.9588                 NA 0.0062   0e+00
IRS1             79 149.0802                 NA 0.0067   0e+00
ADRB2           244 157.1719                 NA 0.0100   1e-04
CAPN10          127 159.8718                 NA 0.0118   1e-04
ABCC8            86 163.9569                 NA 0.0142   1e-04
AKT1             82 166.3992                 NA 0.0138   1e-04
TNF              18 166.8583                 NA 0.0129   1e-04
IAPP             17 187.1393                 NA 0.0267   2e-04
AGER            362 188.5358                 NA 0.0256   3e-04
PAX4            357 194.8055                 NA 0.0306   3e-04
INS             412 201.5425                 NA 0.0378   4e-04
ADRB3           242 202.8109                 NA 0.0379   4e-04
HNF4A           225 204.0128                 NA 0.0390   5e-04
IRS2             80 208.3986                 NA 0.0429   5e-04
STX1A           375 210.6823                 NA 0.0455   6e-04
```

3.2.4 Αποτελέσματα για Υπέρταση.

Εικόνα 3.2.4.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Υπέρταση.



Εικόνα 3.2.4.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Υπέρταση όπου RP.out=RPadvance.

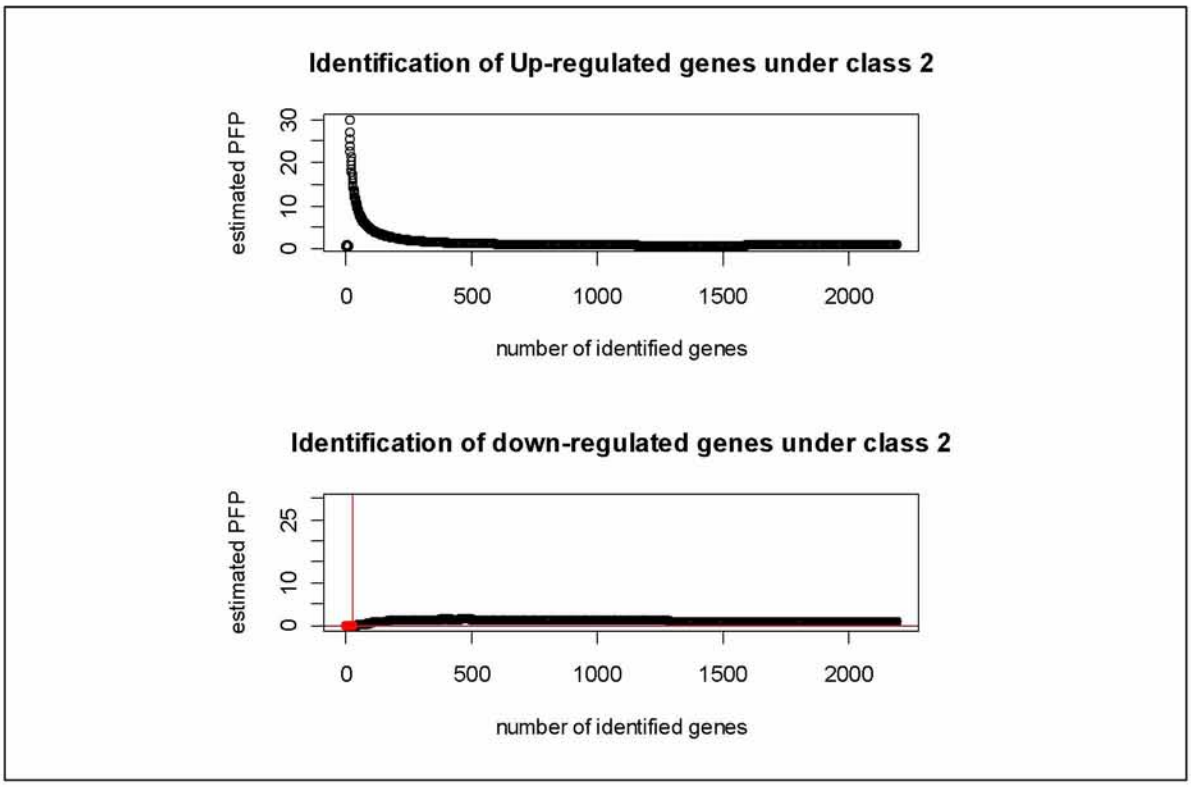


Εικόνα 3.2.4.3: Αποτελέσματα συνάρτησης topGene() για Υπέρταση με cutoff = 0.05 .

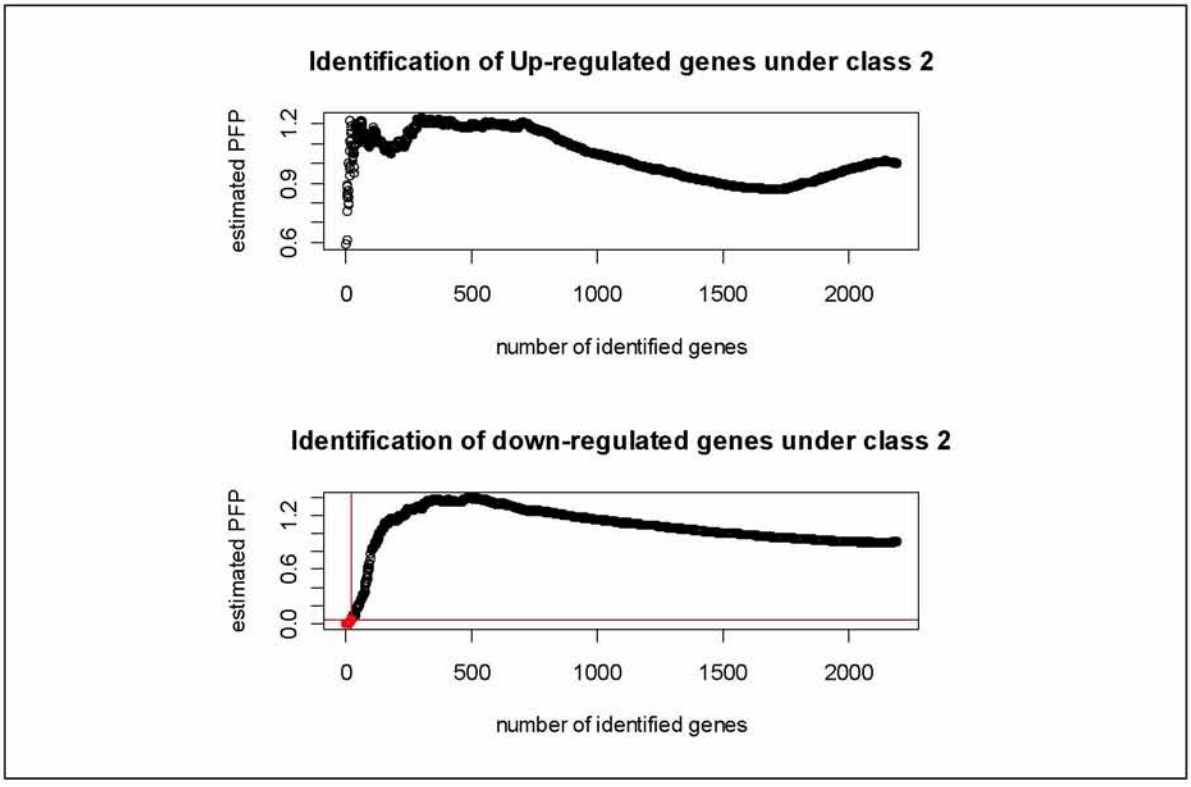
```
Table2: Genes called significant under class1 > class2
$Table1
NULL
$Table2
        gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
AGT            651  42.6582                 NA 0.0000   0e+00
AGTR1          615  80.6428                 NA 0.0000   0e+00
ADD1           611  81.2455                 NA 0.0000   0e+00
EDN1             6  87.0292                 NA 0.0025   0e+00
GNB3           630  87.5318                 NA 0.0020   0e+00
EDNRA            7 107.1224                 NA 0.0033   0e+00
PMP22          371 111.6455                 NA 0.0043   0e+00
BDKRB2         437 119.9064                 NA 0.0038   0e+00
NPPA             4 128.3711                 NA 0.0033   0e+00
SCNN1A         618 129.6051                 NA 0.0030   0e+00
BMPR2            2 132.4847                 NA 0.0027   0e+00
APOE           126 132.6062                 NA 0.0025   0e+00
ADRB2          193 133.2377                 NA 0.0031   0e+00
NOS3            27 136.0990                 NA 0.0036   0e+00
DRD1             9 137.5952                 NA 0.0033   0e+00
PTGIS          335 138.3530                 NA 0.0031   0e+00
NPPC            19 166.2912                 NA 0.0100   1e-04
CYP11B2       1253 169.9836                 NA 0.0117   1e-04
ADM             70 170.3041                 NA 0.0116   1e-04
INSR           573 172.4272                 NA 0.0115   1e-04
ACE            382 174.5190                 NA 0.0114   1e-04
CAT             97 198.4442                 NA 0.0264   3e-04
ESR2           294 199.4744                 NA 0.0257   3e-04
NPR3           209 200.4733                 NA 0.0250   3e-04
NPR1            54 208.6935                 NA 0.0308   4e-04
NR3C2          646 218.9588                 NA 0.0400   6e-04
EDN2           182 220.3810                 NA 0.0396   6e-04
SCNN1G         672 222.7110                 NA 0.0407   6e-04
```

3.2.5 Αποτελέσματα για Σκλήρυνση κατά πλάκας.

Εικόνα 3.2.5.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Σκλήρυνση κατά πλάκα.



Εικόνα 3.2.5.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Σκλήρυνση κατά πλάκα όπου RP.out=RPadvance.

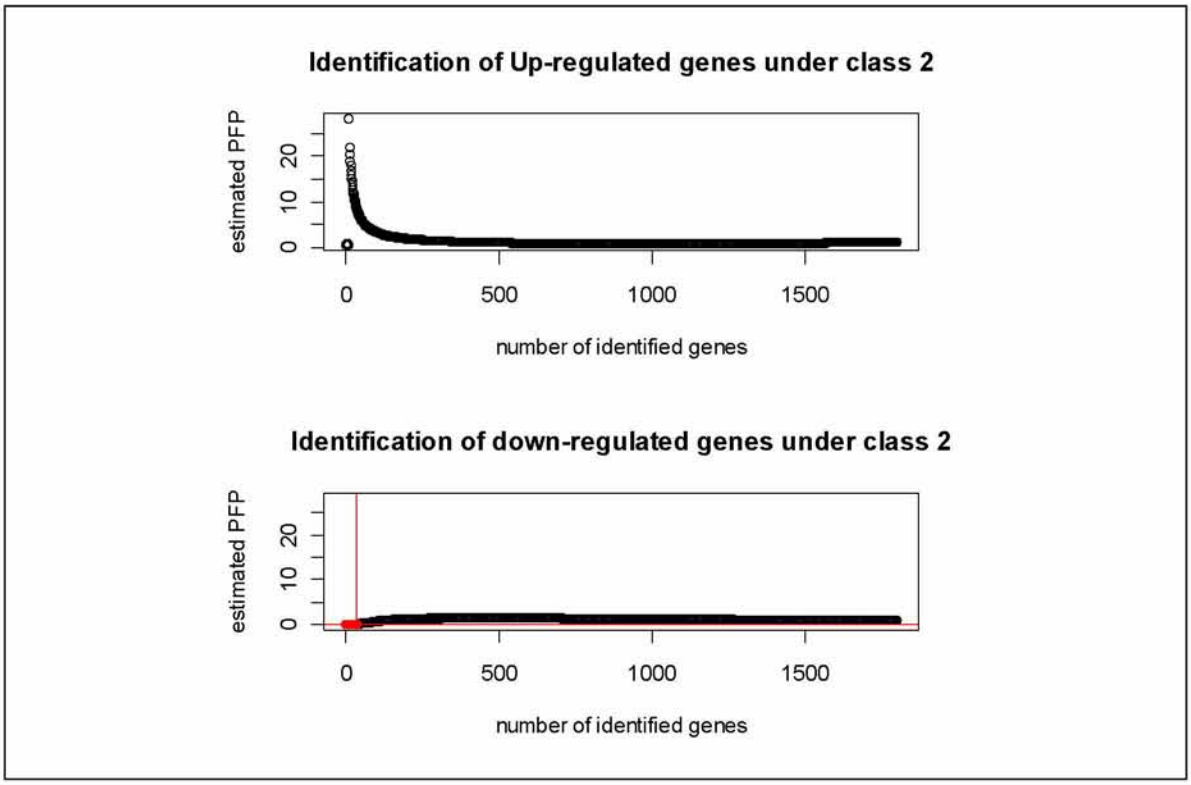


Εικόνα 3.2.5.3: Αποτελέσματα συνάρτησης topGene() για Σκλήρυνση κατά πλάκας με cutoff = 0.05 .

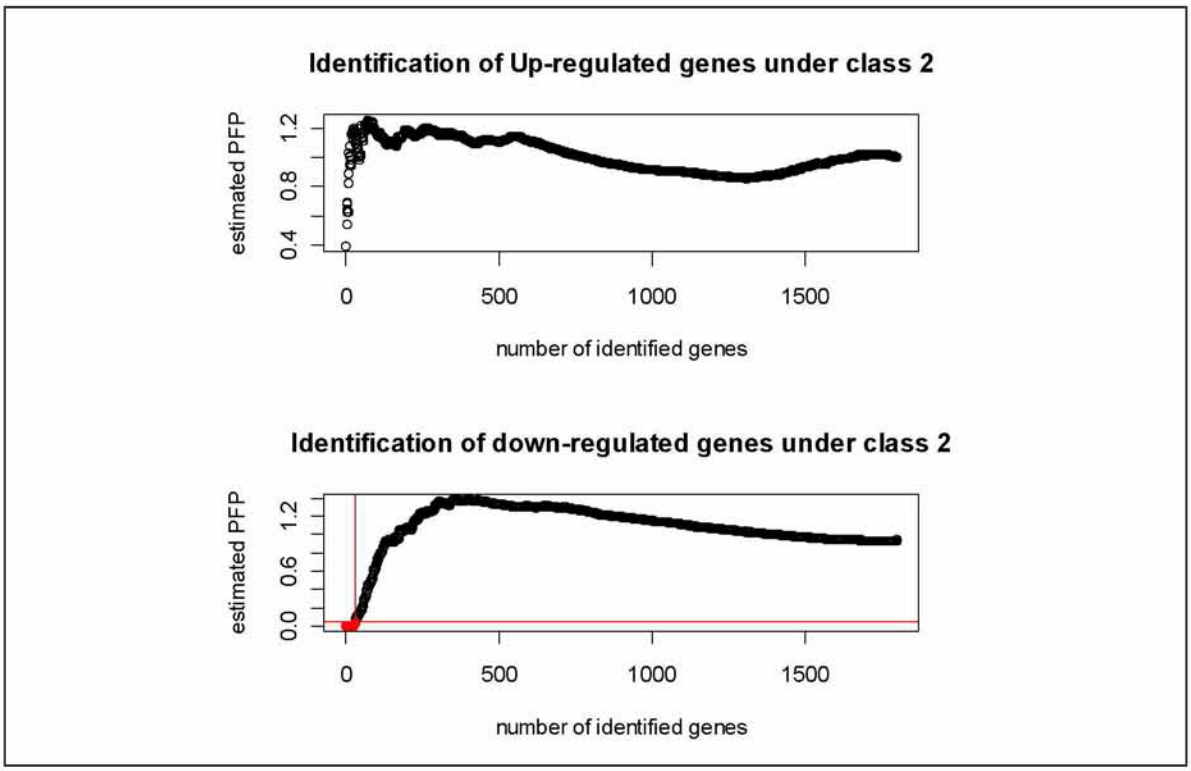
```
Table2: Genes called significant under class1 > class2
$Table1
NULL
$Table2
         gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
PTPRC            37  40.3821                 NA 0.0000   0e+00
MBP               3  66.9317                 NA 0.0000   0e+00
APOE             18  77.8179                 NA 0.0000   0e+00
IL6               7  90.7758                 NA 0.0000   0e+00
PRKCA           758 108.1401                 NA 0.0000   0e+00
CD24            759 108.9801                 NA 0.0000   0e+00
CTLA4           192 113.1517                 NA 0.0000   0e+00
VDR              90 115.2253                 NA 0.0000   0e+00
CRYAB           546 125.0704                 NA 0.0011   0e+00
IL4              20 138.7308                 NA 0.0010   0e+00
TNF               6 149.4443                 NA 0.0018   0e+00
IL1B              9 220.4041                 NA 0.0175   1e-04
IL10             10 222.7006                 NA 0.0185   1e-04
IFNG             11 233.4685                 NA 0.0229   1e-04
TNFRSF1A         12 239.0174                 NA 0.0280   2e-04
WT1            1424 241.7127                 NA 0.0294   2e-04
SH2D2A          768 243.9157                 NA 0.0294   2e-04
SPP1            481 252.9871                 NA 0.0361   3e-04
MOG              41 256.7527                 NA 0.0379   3e-04
CREBBP          877 257.3165                 NA 0.0365   3e-04
ALDH6A1         760 266.3026                 NA 0.0419   4e-04
MYO15A          784 271.0971                 NA 0.0445   4e-04
```

3.2.6 Αποτελέσματα για Παχυσαρκία.

Εικόνα 3.2.6.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Παχύσαρκία.



Εικόνα 3.2.6.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Παχυσαρκία όπου RP.out=RPadvance.



Εικόνα 3.2.6.3: Αποτελέσματα συνάρτησης topGene() για Παχυσαρκία με cutoff = 0.05 .

```
Table2: Genes called significant under class1 > class2

$Table1
NULL

$Table2
         gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
UCP2             53  25.5365                 NA 0.0000   0e+00
PPARG            98  40.7176                 NA 0.0000   0e+00
LEP               4  41.7575                 NA 0.0000   0e+00
UCP3             52  44.1274                 NA 0.0000   0e+00
ADRB2           139  44.1569                 NA 0.0000   0e+00
MC4R             39  50.1235                 NA 0.0000   0e+00
LPL               5  56.0249                 NA 0.0000   0e+00
LEPR            560  59.3184                 NA 0.0000   0e+00
POMC             28  64.5295                 NA 0.0000   0e+00
TNF              11  68.2749                 NA 0.0000   0e+00
PCSK1           594  72.6045                 NA 0.0000   0e+00
FTO               2  92.2571                 NA 0.0000   0e+00
MC3R            232  92.5109                 NA 0.0000   0e+00
LIPE            155  93.2975                 NA 0.0000   0e+00
ADIPOQ           38  94.6387                 NA 0.0000   0e+00
NR3C1            85 101.3787                 NA 0.0000   0e+00
RETN             37 108.9171                 NA 0.0000   0e+00
AGRP             35 116.8727                 NA 0.0006   0e+00
IRS2            178 117.3269                 NA 0.0005   0e+00
NR0B2           577 126.2657                 NA 0.0005   0e+00
GNB3           1193 127.1128                 NA 0.0005   0e+00
ADRB3            43 143.4939                 NA 0.0014   0e+00
NPY              33 151.5996                 NA 0.0013   0e+00
SORBS1          563 154.2897                 NA 0.0033   0e+00
GHRL             92 164.4012                 NA 0.0048   1e-04
UCP1             31 172.6660                 NA 0.0077   1e-04
PPARD            99 177.6343                 NA 0.0089   1e-04
HSD11B1         140 193.6671                 NA 0.0179   3e-04
```

3.3 Αποτελέσματα R-PROJECT για συσχετισμό δύο ασθενειών.

Στα επόμενα αποτελέσματα που ακολουθούν χρησιμοποιήσαμε μία ειδική συνθήκη που παρέχει το πακέτο RankProd, την στατιστική ανάλυση των αποτελεσμάτων δύο ασθενειών όπου θέτει την μία ενάντια στην άλλη, σκοράροντας τα αποτελέσματα και παρουσιάζοντας τα διαφορικά εκφρασμένα γονίδια που βρέθηκε να σχετίζονται και με τις δύο ασθένειες.

Η διαδικασία αυτή εκτελείται πολύ απλά, θέτοντας τα γονίδια της μίας ασθένειας μαζί με τις σχετικές τους πληροφορίες να ανήκουν στην κλάση 1 και τα γονίδια της άλλης ασθένειες μαζί με τις σχετικές τους πληροφορίες να ανήκουν στην κλάση 2.

Όπως και πιο πάνω τα αποτελέσματα παρουσιάζονται μέσα από τις τρείς βασικές αναλύσεις που μας προσφέρει το πακέτο RankProd που είναι:

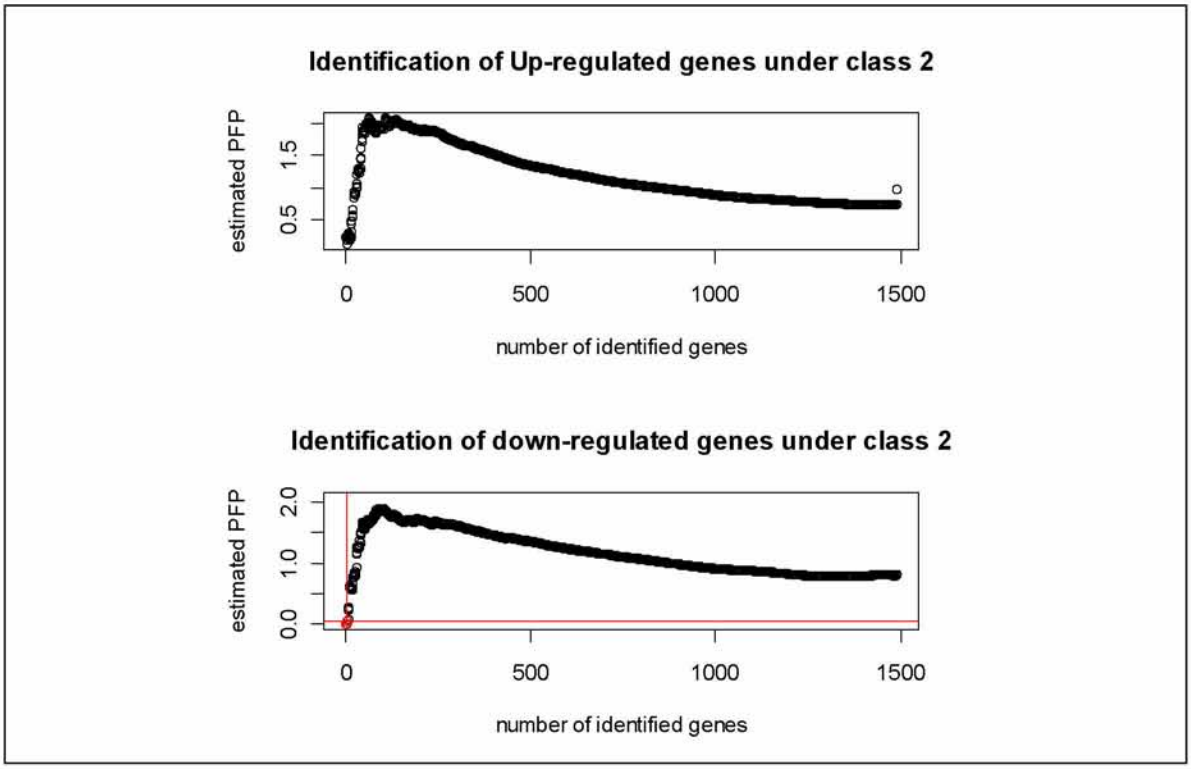
-Παρουσίαση των αποτελεσμάτων σε γραφική παράσταση με βάση την συνάρτηση «plotRP» με δεδομένα εισόδου RP.out και ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05. (cutoff = 0.05)

-Παρουσίαση των αποτελεσμάτων σε γραφική παράσταση με βάση την συνάρτηση «plotRP» με δεδομένα εισόδου RP.out=RPadvance και ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05. (cutoff = 0.05)

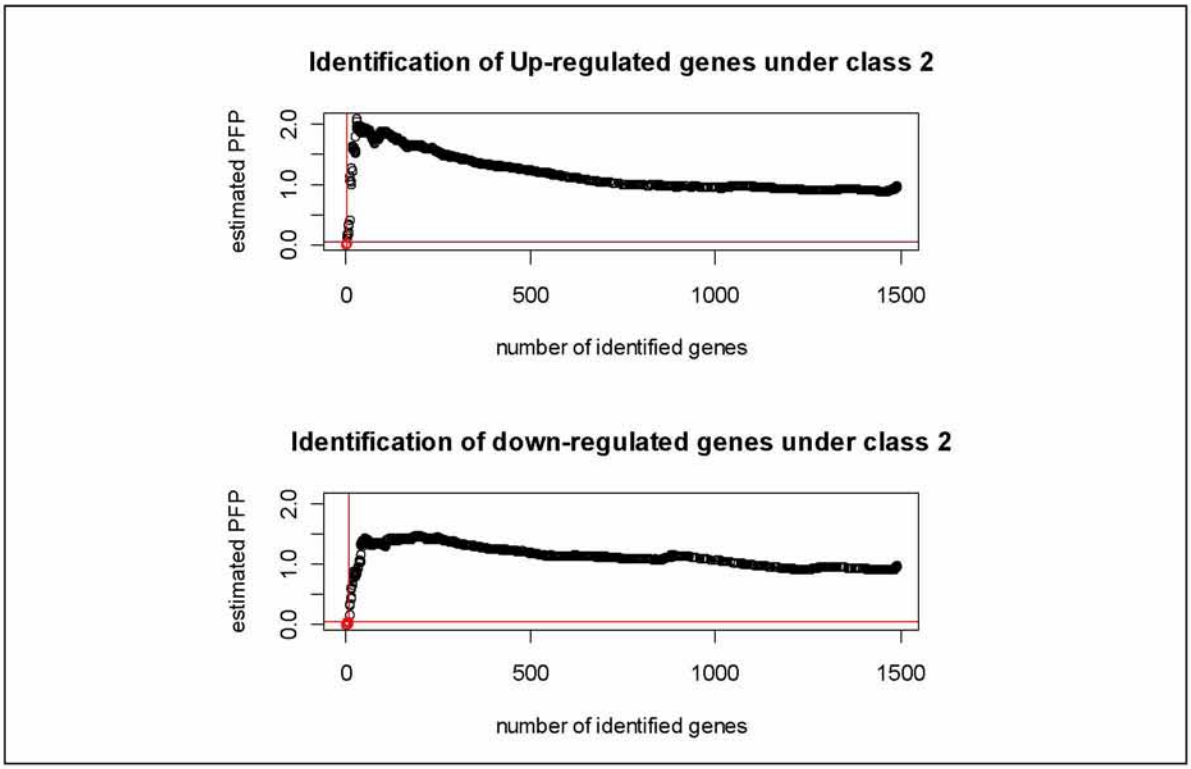
-Παρουσίαση των υψηλά υποψήφιων γονιδίων που σχετίζονται με μία ασθένεια με την συνάρτηση «topGene» με ποσοστό ψευδώς θετικών προβλέψεων (pfp) να είναι μικρότερο του 0.05. (cutoff = 0.05). Στο πρώτο πίνακα παρουσιάζονται τα υψηλά διαφορετικά εκφρασμένα γονίδια της μίας ασθένειας, ενώ στον δεύτερο πίνακα παρουσιάζονται τα υψηλά διαφορετικά εκφρασμένα γονίδια της άλλης ασθένειας .

3.3.1 Αποτελέσματα Διαβήτη τύπου Ι ενάντια σε Διαβήτη τύπου ΙΙ.

Εικόνα 3.3.1.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Διαβήτη Ι.



Εικόνα 3.3.1.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Διαβήτη Ι όπου RP.out=RPadvance.

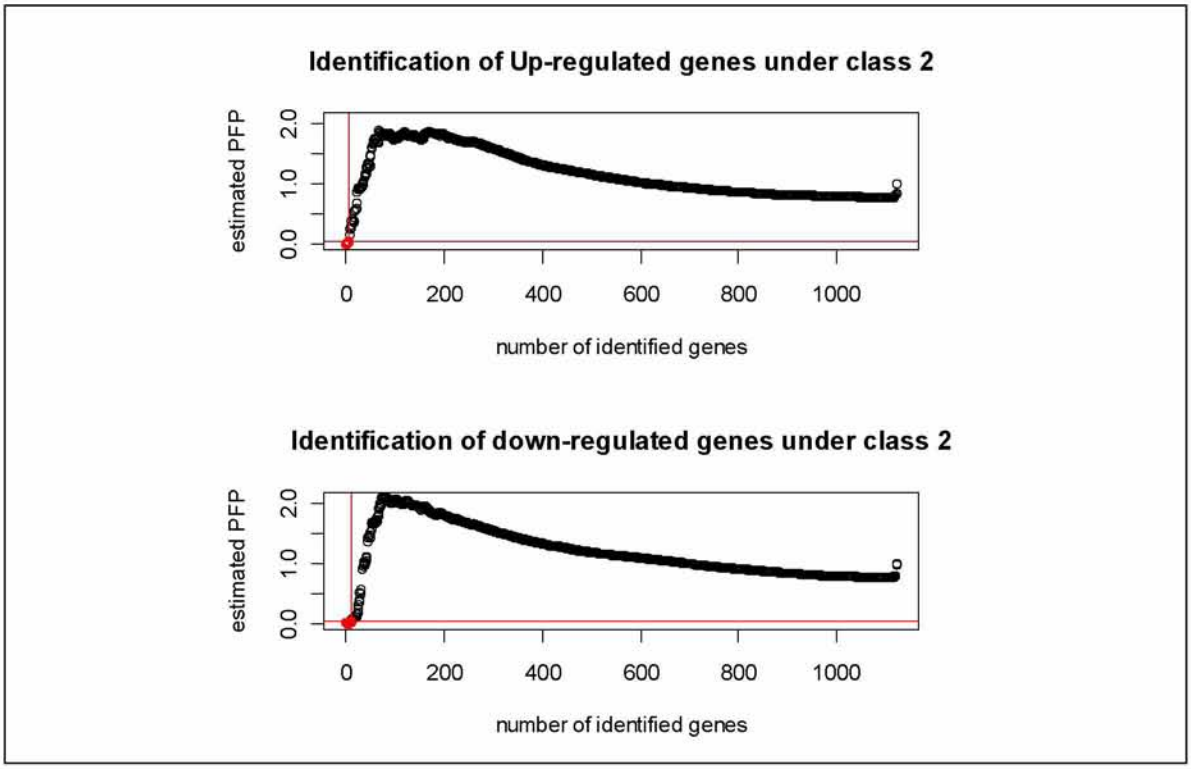


Εικόνα 3.3.1.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου Ι ενάντια με Διαβήτη τύπου ΙΙ με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου Ι ενώ στον πίνακα 2 δίνεται βαρύτητα στα γονίδα του με Διαβήτη τύπου ΙΙ.

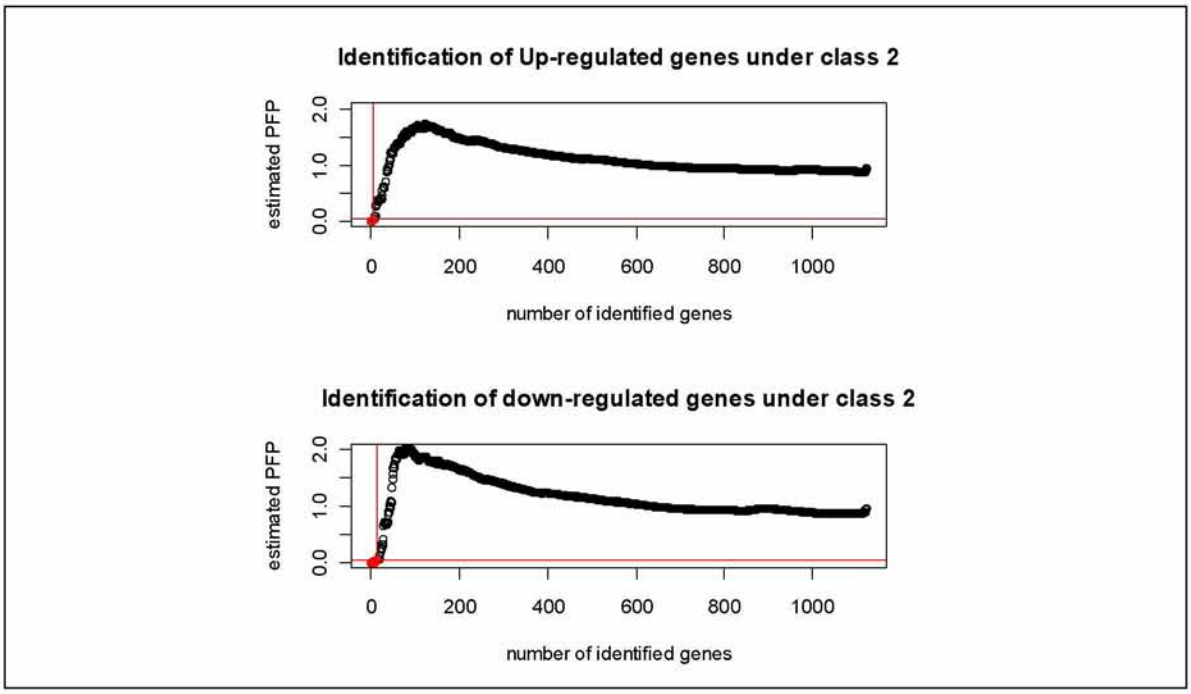
```
Table1: Genes called significant under class1 < class2

Table2: Genes called significant under class1 > class2

$Table1
      gene.index  RP/Rsum FC:(class1/class2)  pfp P.value
B2M          254 107.2820                  0 0.02       0
THBS1        226 127.5237                  0 0.03       0

$Table2
        gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
SCP2           252  35.6341                Inf 0.0000   0e+00
SH2B1          349  82.0732                Inf 0.0000   0e+00
EGF            155  97.4594                Inf 0.0033   0e+00
NPPB           207 103.8082                Inf 0.0050   0e+00
EDNRA           15 146.2588                Inf 0.0360   1e-04
CPT1A          359 148.1610                Inf 0.0317   1e-04
TNFSF13        294 149.1831                Inf 0.0286   1e-04
BCL2           233 157.0771                Inf 0.0300   2e-04
IL1A            27 164.8626                Inf 0.0378   2e-04
```

3.3.2 Αποτελέσματα Διαβήτη τύπου Ι ενάντια σε Υπέρταση.

Εικόνα 3.3.2.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Υπέρταση.



Εικόνα 3.3.2.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Υπέρταση όπου RP.out=RPadvance.

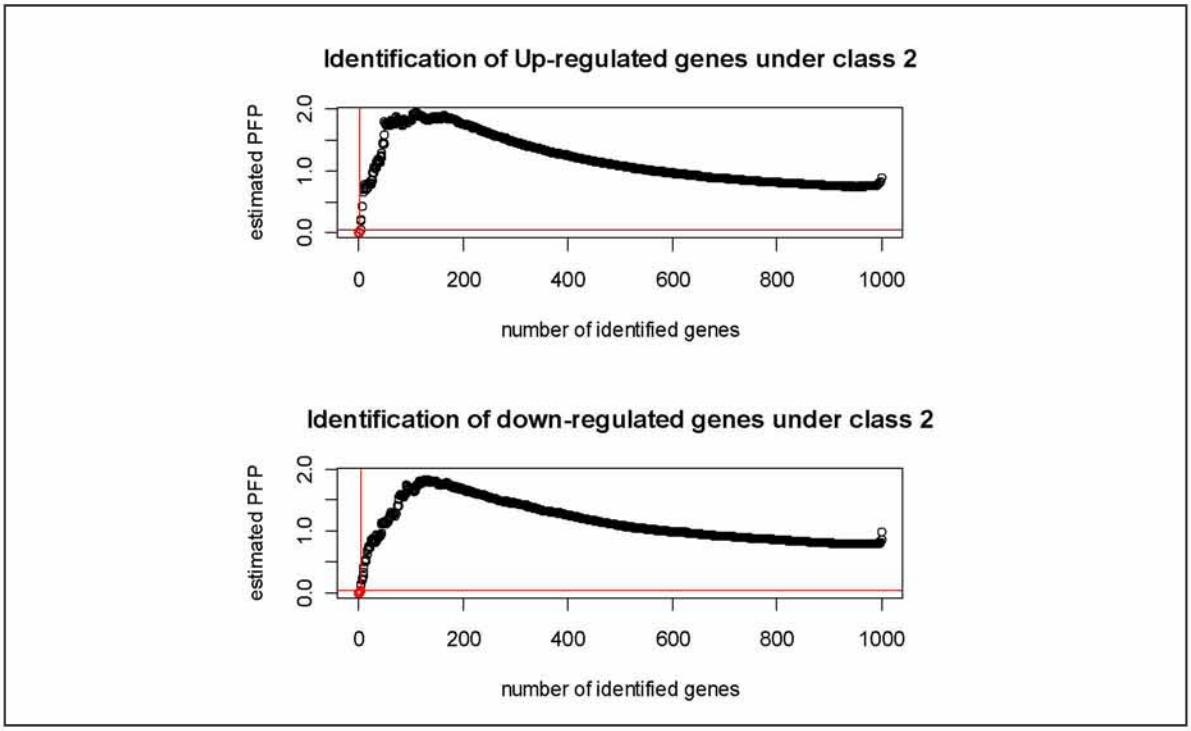


Εικόνα 3.3.2.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου Ι ενάντια με Υπέρταση, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου Ι ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Υπέρτασης.

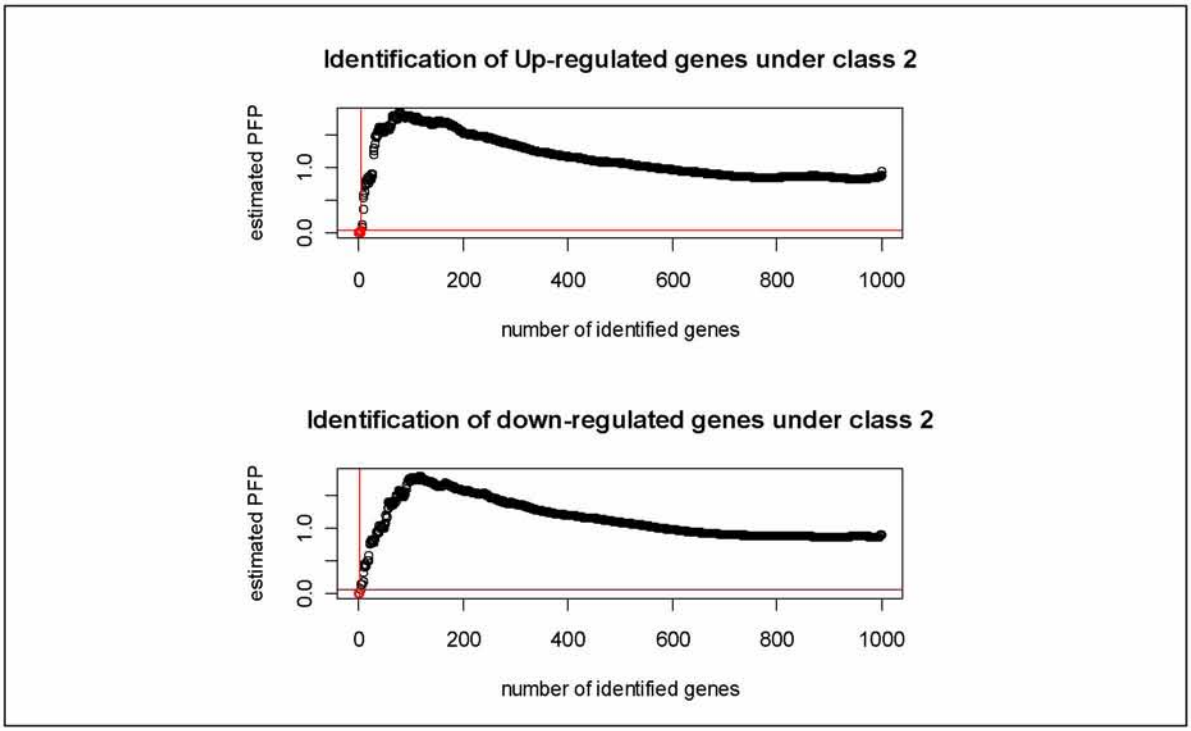
```
Table1: Genes called significant under class1 < class2

Table2: Genes called significant under class1 > class2

$Table1
      gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
CTLA4         54  23.7782                  0 0.0000   0e+00
IRS2          39  76.0964                  0 0.0100   0e+00
VDR           64  88.8702                  0 0.0100   0e+00
HNF1A        191  90.0148                  0 0.0075   0e+00
ICAM1         35 110.1650                  0 0.0260   1e-04
IRS1          38 123.5167                  0 0.0350   2e-04
IL6           23 125.2634                  0 0.0329   2e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
ADD1          376  42.6497                Inf 0.0000   0e+00
BDKRB2        786  52.2650                Inf 0.0000   0e+00
AGT           372  53.3849                Inf 0.0000   0e+00
AGTR1         710  81.7584                Inf 0.0025   0e+00
GNB3          948  88.3883                Inf 0.0040   0e+00
PMP22         982  90.9491                Inf 0.0033   0e+00
ADRB2         738 107.3933                Inf 0.0129   1e-04
EDN1           10 107.6076                Inf 0.0112   1e-04
ESR2          872 114.3943                Inf 0.0144   1e-04
NPPA           88 131.4367                Inf 0.0330   3e-04
EDNRA          13 131.9058                Inf 0.0300   3e-04
NPPC          203 137.2982                Inf 0.0358   4e-04
F2            951 140.9751                Inf 0.0392   5e-04
NPR3          205 142.8614                Inf 0.0407   5e-04
NPR1          204 143.4226                Inf 0.0387   5e-04
```

3.3.3 Αποτελέσματα Διαβήτη τύπου Ι ενάντια σε Σκλήρυνση κατά πλάκας.

Εικόνα 3.3.3.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Σκλήρυνση κατά πλάκας.



Εικόνα 3.3.3.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Σκλήρηνση κατά πλάκας όπου RP.out=RPadvance.

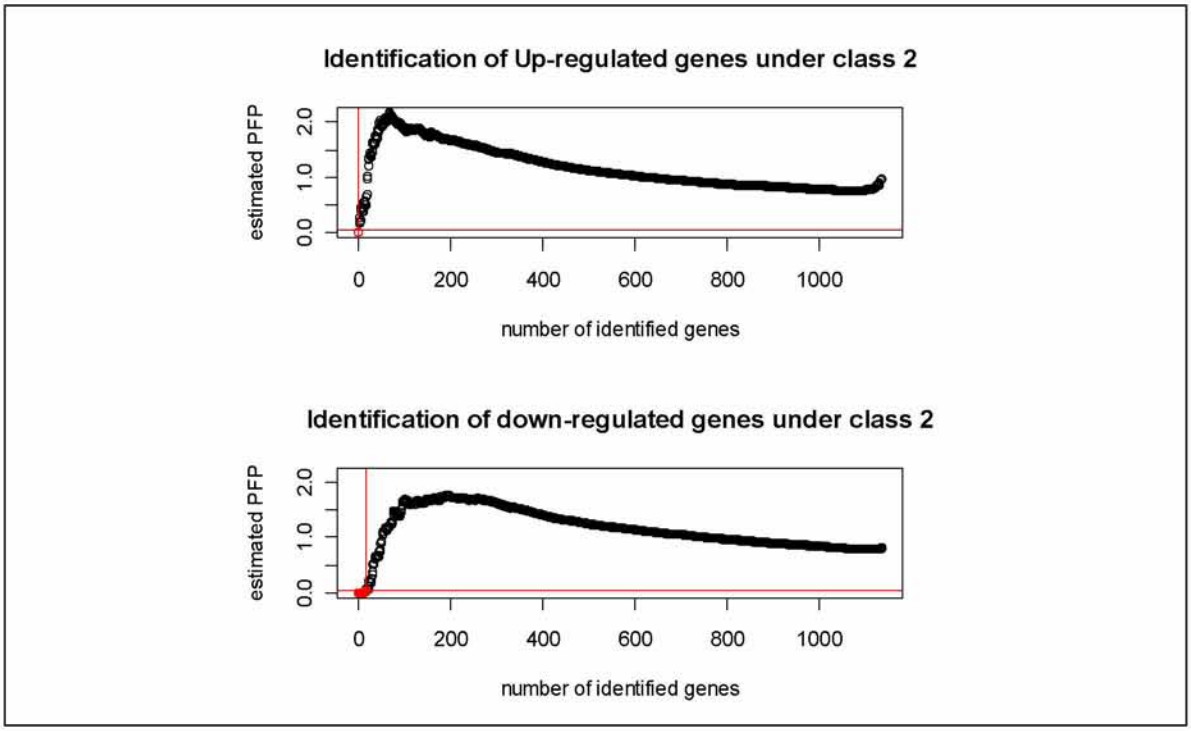


Εικόνα 3.3.3.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου Ι ενάντια με Σκλήρηνση κατά πλάκας, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου Ι ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Σκλήρυνσης κατά πλάκας.

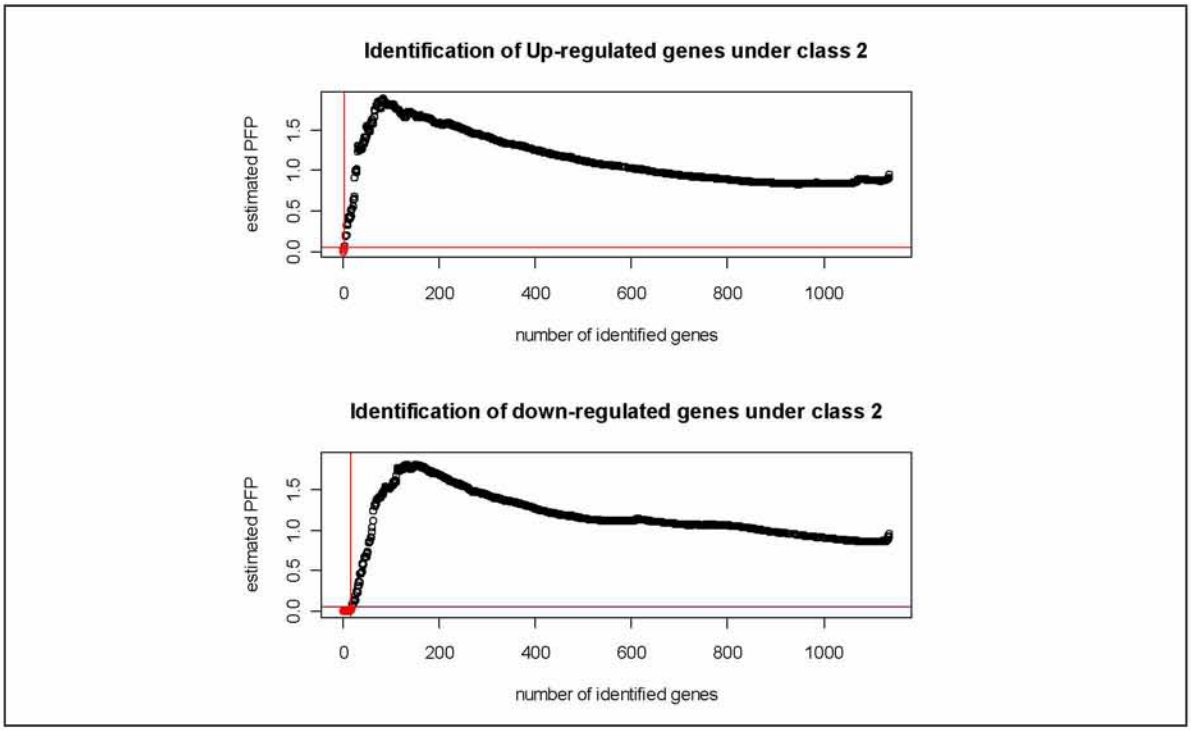
```
Table1: Genes called significant under class1 < class2

Table2: Genes called significant under class1 > class2

$Table1
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
IRS2           35  54.8805                  0 0.0000   0e+00
SUMO4         519  65.0662                  0 0.0000   0e+00
IRS1           34  73.4723                  0 0.0067   0e+00
TAP2          424  96.8889                  0 0.0150   1e-04
PTPN22        151 103.1821                  0 0.0340   2e-04
GCK           201 109.4969                  0 0.0400   2e-04

$Table2
      gene.index RP/Rsum FC:(class1/class2)    pfp P.value
PTPRC         54 56.8961                Inf 0.0000       0
MBP           42 66.4146                Inf 0.0000       0
PRKCA        552 88.6158                Inf 0.0067       0
```

3.3.4 Αποτελέσματα Διαβήτη τύπου Ι ενάντια σε Παχυσαρκία.

Εικόνα 3.3.4.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Παχυσαρκία.



Εικόνα 3.3.4.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη Ι ενάντια με Παχυσαρκία όπου RP.out=RPadvance.

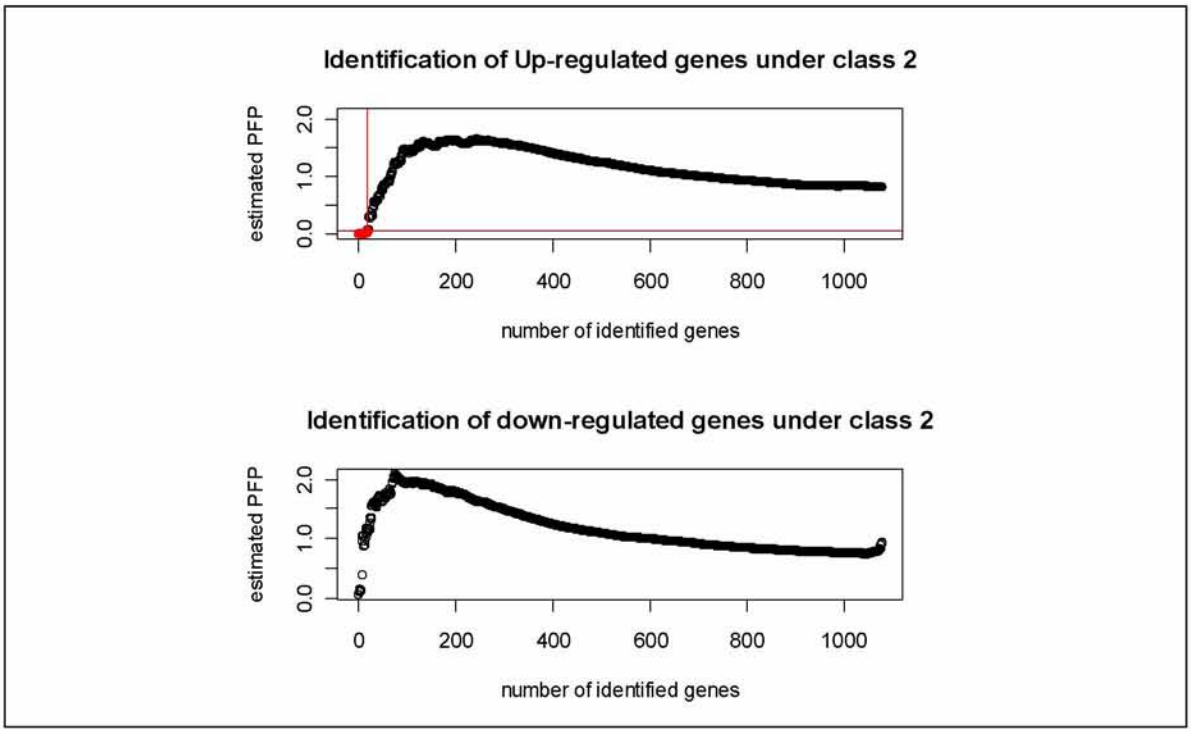


Εικόνα 3.3.4.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου Ι ενάντια με Παχυσαρκία, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου Ι ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Παχυσαρκίας.

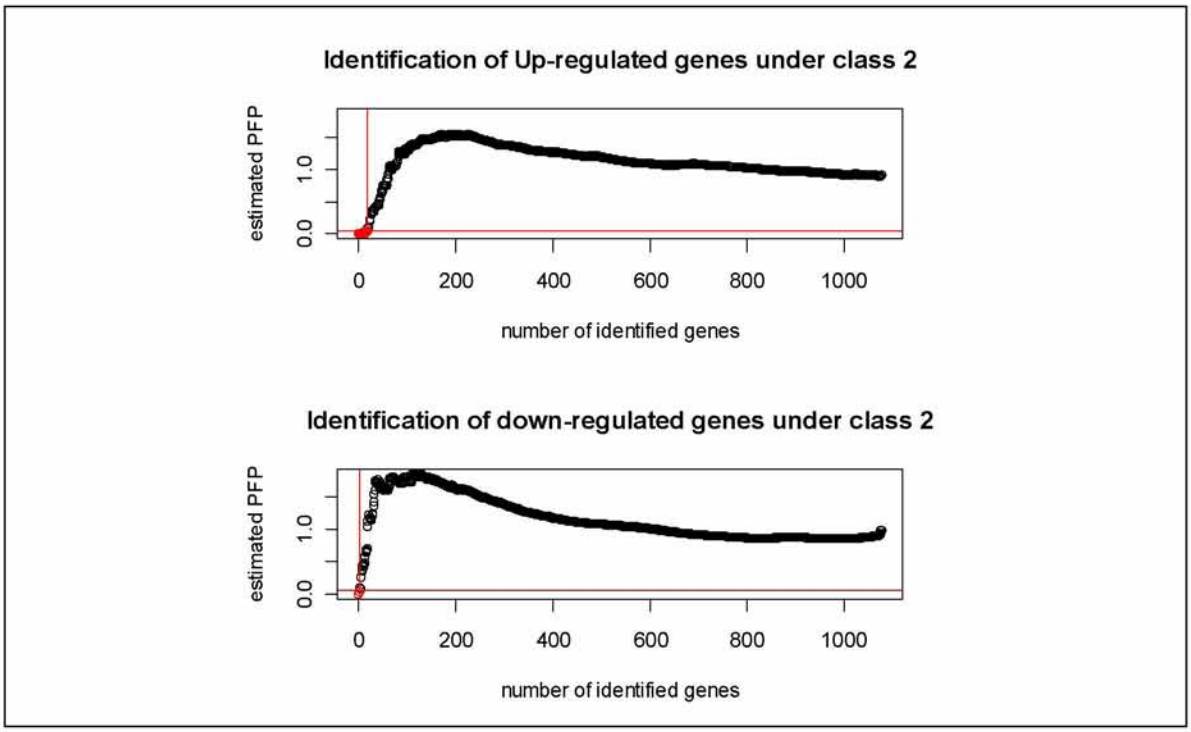
```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2

$Table1
        gene.index  RP/Rsum FC:(class1/class2)   pfp P.value
PTPN22         157  45.1497                  0 0.000   0e+00
NEUROD1        136  96.6674                  0 0.015   0e+00
VDR             67 103.1377                  0 0.030   1e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
MC4R          336  34.4202                Inf 0.0000   0e+00
LIPE          373  44.3919                Inf 0.0000   0e+00
UCP3          415  46.2578                Inf 0.0000   0e+00
ADRB2         748  47.4013                Inf 0.0000   0e+00
UCP2          413  59.8925                Inf 0.0000   0e+00
PCSK1        1020  63.3182                Inf 0.0000   0e+00
AGRP          174  73.5345                Inf 0.0000   0e+00
NR3C1         423  74.5650                Inf 0.0000   0e+00
LEP            75  78.5986                Inf 0.0000   0e+00
LEPR          819  82.3356                Inf 0.0000   0e+00
POMC           89  86.6300                Inf 0.0009   0e+00
SORBS1        927  89.4212                Inf 0.0025   0e+00
PPARG         160 120.6530                Inf 0.0192   2e-04
PPARD         208 123.0689                Inf 0.0200   2e-04
GNB3          969 126.1284                Inf 0.0200   3e-04
FTO           958 129.4565                Inf 0.0200   3e-04
SREBF1        392 135.4908                Inf 0.0300   4e-04
LPL           330 137.2287                Inf 0.0306   5e-04
```

3.3.5 Αποτελέσματα Διαβήτη τύπου ΙΙ ενάντια σε Παχυσαρκία.

Εικόνα 3.3.5.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Παχυσαρκία.



Εικόνα 3.3.5.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Παχυσαρκία όπου RP.out=RPadvance.

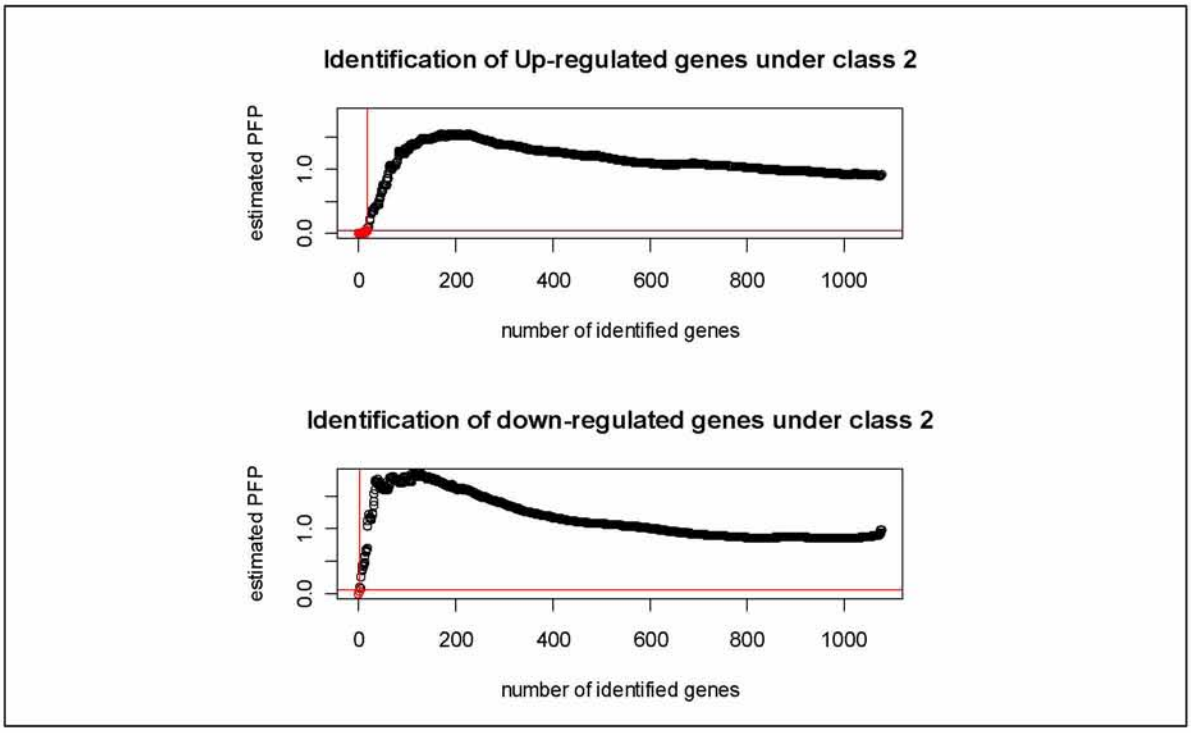


Εικόνα 3.3.5.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου ΙΙ ενάντια με Παχυσαρκία, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου ΙΙ ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Παχυσαρκίας

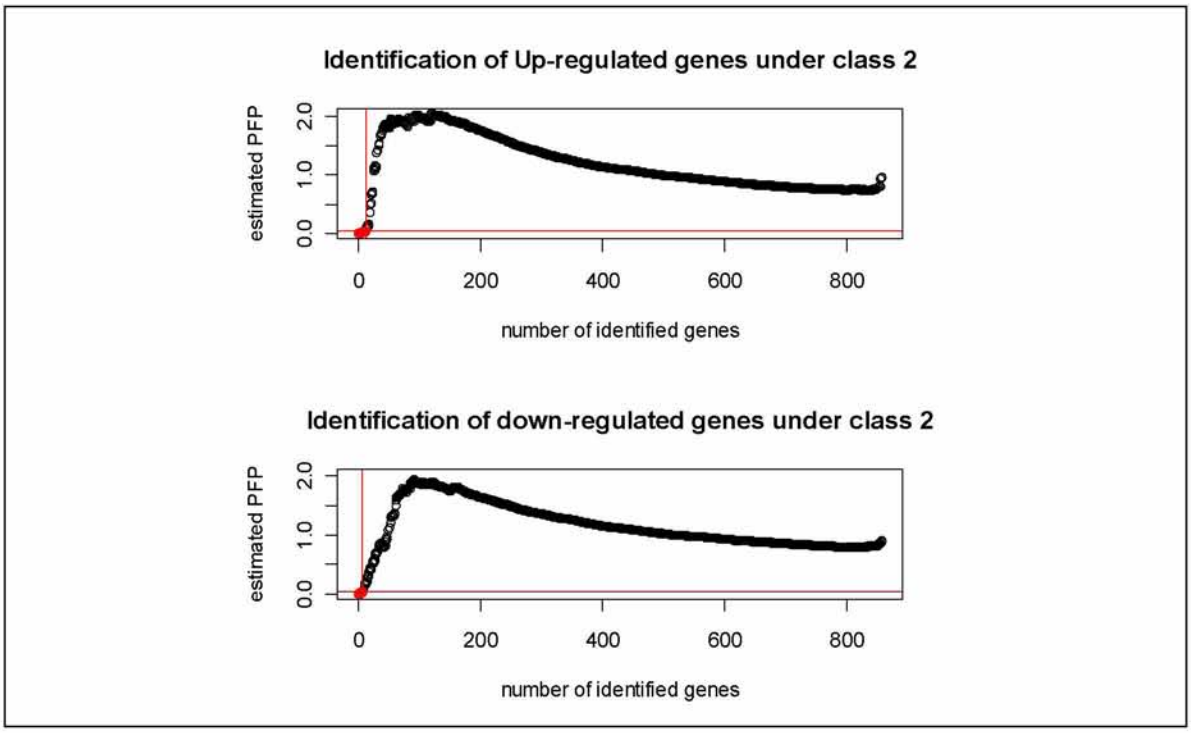
```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2

$Table1
        gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
UCP2            48  27.3093                  0 0.0000   0e+00
LIPE           130  44.3818                  0 0.0000   0e+00
LEPR           368  57.8174                  0 0.0000   0e+00
PCSK1          378  61.9189                  0 0.0000   0e+00
POMC            25  63.1686                  0 0.0000   0e+00
NPY             30  71.4930                  0 0.0000   0e+00
NR3C1           75  77.1131                  0 0.0000   0e+00
LPL              3  89.8765                  0 0.0075   1e-04
FTO              1  92.1624                  0 0.0089   1e-04
GNB3           880  99.9234                  0 0.0110   1e-04
LEP              2 106.0942                  0 0.0127   1e-04
GHRL            81 119.7152                  0 0.0283   3e-04
HSD11B1        119 120.6240                  0 0.0292   4e-04
ADRB2          118 121.7609                  0 0.0293   4e-04
PPARD           87 121.8970                  0 0.0273   4e-04
SORBS1         370 136.1198                  0 0.0438   6e-04
SREBF1          78 137.5820                  0 0.0447   7e-04
NR1H3          201 137.6548                  0 0.0422   7e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
BCHE          983  76.7207                Inf 0.0000   0e+00
AGER          998 107.1838                Inf 0.0450   1e-04
KCNJ11        106 110.5227                Inf 0.0467   1e-04
```

3.3.6 Αποτελέσματα Διαβήτη τύπου ΙΙ ενάντια σε Σκλήρυνση κατά πλάκας.

Εικόνα 3.3.6.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Σκλήρυνση κατά πλάκα.



Εικόνα 3.3.6.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Σκλήρησης κατά πλάκα όπου RP.out=RPadvance.

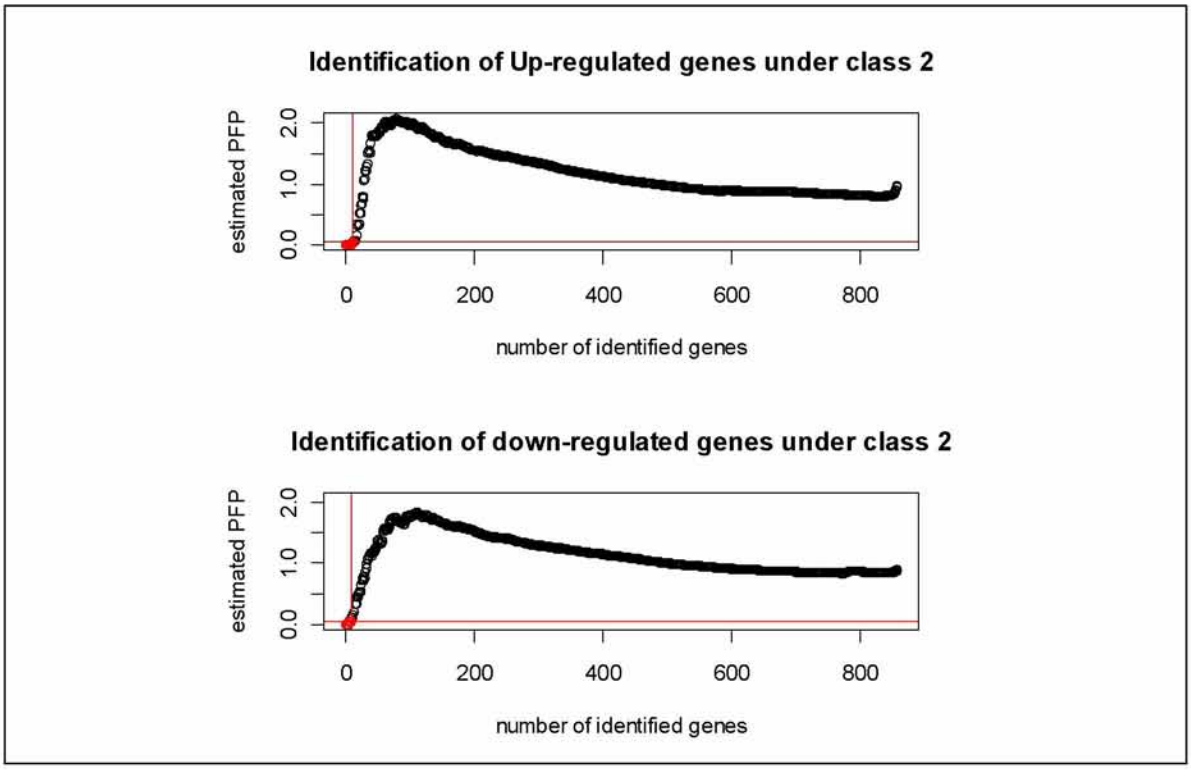


Εικόνα 3.3.6.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου II ενάντια με Σκλήρηνση κατά πλάκας, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου II ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Σκλήρυνσης κατά πλάκας.

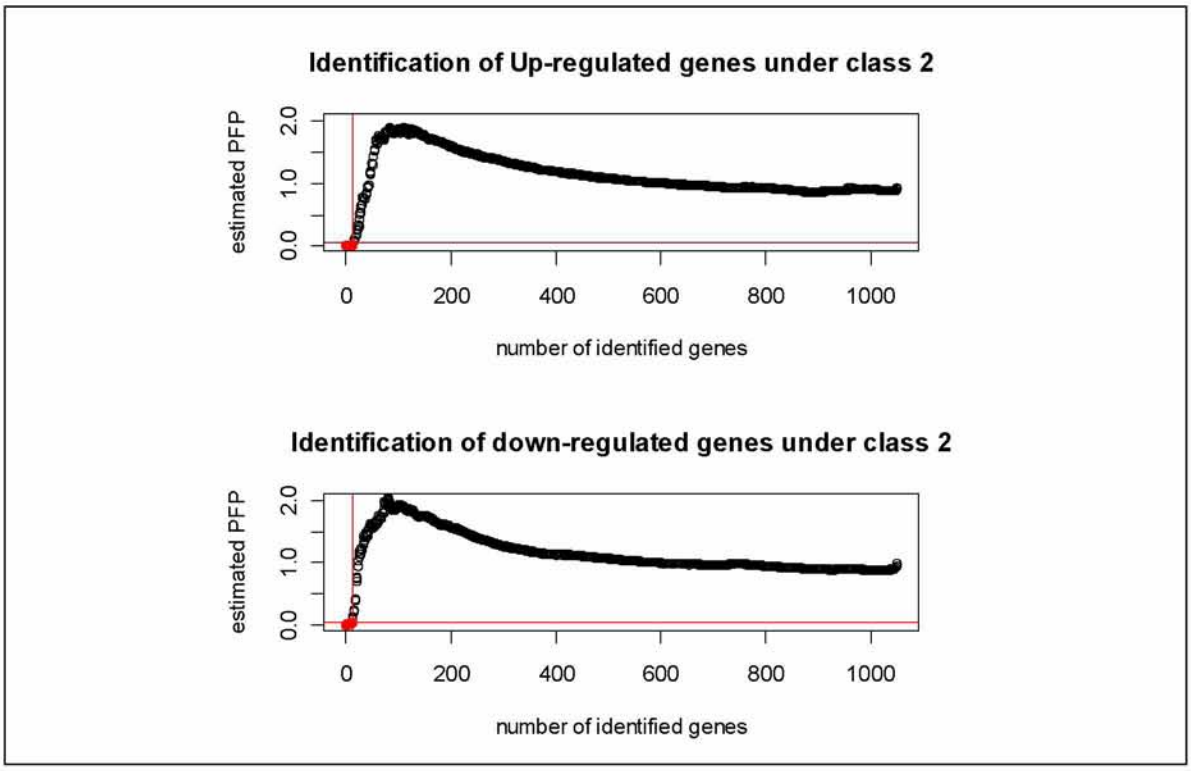
```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2

$Table1
         gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
PPARGC1A        120  55.1454                  0 0.0000   0e+00
RETN             51  60.8311                  0 0.0050   0e+00
GCK               1  63.5400                  0 0.0033   0e+00
KCNJ11           73  64.5608                  0 0.0025   0e+00
ADRB3           204  74.5023                  0 0.0020   0e+00
IRS2             68  76.6502                  0 0.0017   0e+00
TCF7L2           59  82.1998                  0 0.0057   0e+00
BCHE            231  84.5216                  0 0.0088   1e-04
ABCC8            74  98.3031                  0 0.0289   3e-04
HNF4A           192 100.7073                  0 0.0300   3e-04
ADRB2           205 101.8287                  0 0.0282   4e-04
PPARG           103 110.7998                  0 0.0442   6e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
PTPRC         362  32.4857                Inf 0.0000   0e+00
MBP           306  34.4481                Inf 0.0000   0e+00
CTLA4         230  62.3029                Inf 0.0033   0e+00
PRKCA         315  83.1746                Inf 0.0075   0e+00
APOE           49  98.5812                Inf 0.0380   2e-04
VDR           141 105.6009                Inf 0.0517   4e-04
SH2D2A        605 105.9488                Inf 0.0471   4e-04
IL4            12 107.1378                Inf 0.0462   4e-04
```

3.3.7 Αποτελέσματα Διαβήτη τύπου ΙΙ ενάντια σε Υπέρταση.

Εικόνα 3.3.7.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Υπέρταση.



Εικόνα 3.3.7.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Διαβήτη ΙΙ ενάντια με Υπέρταση όπου RP.out=RPadvance.

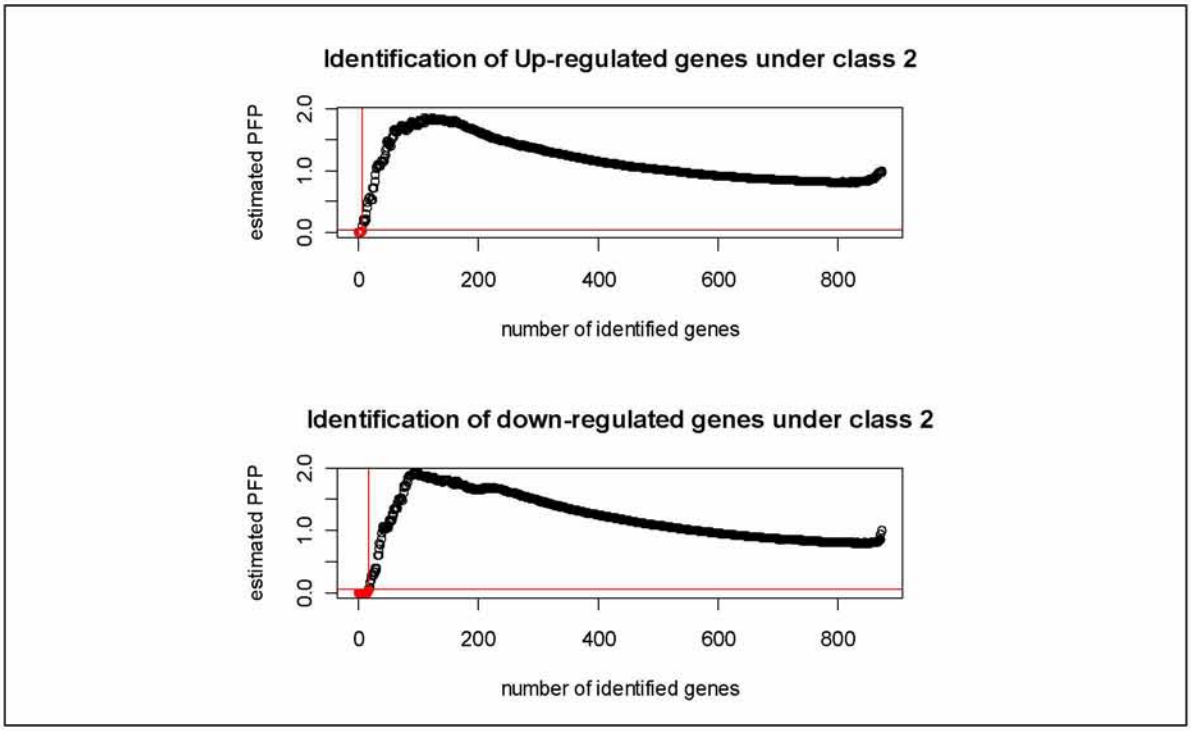


Εικόνα 3.3.7.3: Αποτελέσματα συνάρτησης topGene() για Διαβήτη τύπου II ενάντια με Υπέρταση, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια του Διαβήτη τύπου II ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Υπέρτασης.

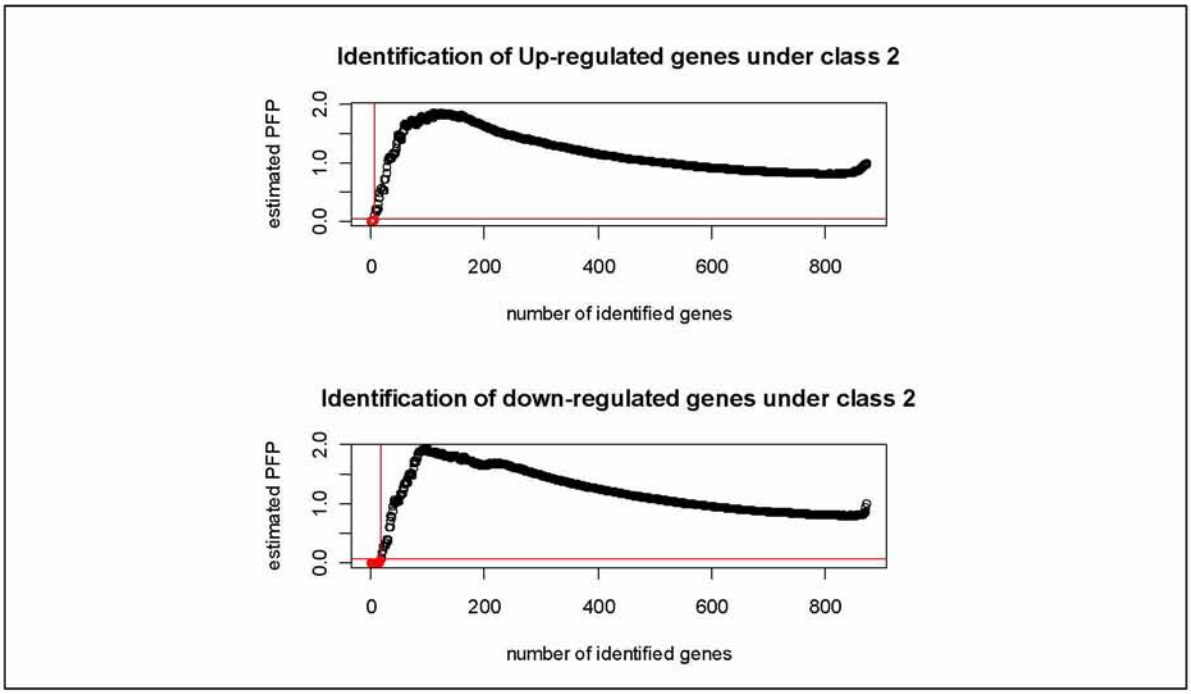
```
Table1: Genes called significant under class1 < class2

Table2: Genes called significant under class1 > class2

$Table1
        gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
AGT            401  45.5660                  0 0.0000   0e+00
BDKRB2         301  51.6454                  0 0.0000   0e+00
GNB3           392  72.7003                  0 0.0000   0e+00
EDN1             5  77.1060                  0 0.0025   0e+00
CYP11B2        882  84.3893                  0 0.0060   0e+00
AGTR1          385  84.4914                  0 0.0050   0e+00
PMP22          266  85.4557                  0 0.0043   0e+00
NPPA             3  85.6321                  0 0.0038   0e+00
NPPC            14  87.9924                  0 0.0033   0e+00
EDNRA            6  93.7739                  0 0.0060   1e-04
ESR2           216 103.3579                  0 0.0118   1e-04
EDN2           143 106.5928                  0 0.0150   2e-04
ADRB2          147 120.1145                  0 0.0292   4e-04
BMPR2            1 125.3993                  0 0.0364   5e-04

$Table2
         gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
GCK             318  79.5968                Inf 0.0000   0e+00
BCHE            365  89.0809                Inf 0.0100   0e+00
KCNJ11           46  89.3388                Inf 0.0067   0e+00
PPARGC1A        887  98.4388                Inf 0.0200   1e-04
CAPN10          229 105.6773                Inf 0.0260   1e-04
AGER            960 106.3368                Inf 0.0233   1e-04
HNF4A           320 106.9643                Inf 0.0200   1e-04
TCF7L2          312 107.8199                Inf 0.0188   1e-04
HNF1A           879 107.9711                Inf 0.0167   1e-04
IRS2            326 109.4971                Inf 0.0160   2e-04
GCKR            888 124.8580                Inf 0.0355   4e-04
STX1A           750 131.7834                Inf 0.0500   6e-04
```

3.3.8 Αποτελέσματα Σκλήρυνσης κατά πλάκας ενάντια σε Παχυσαρκία.

Εικόνα 3.3.8.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Σκλήρυνση κατά Πλάκας ενάντια με Παχυσαρκία.



Εικόνα 3.3.8.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Σκλήρηνση κατά πλάκας ενάντια με Παχυσαρκία όπου RP.out=RPadvance .

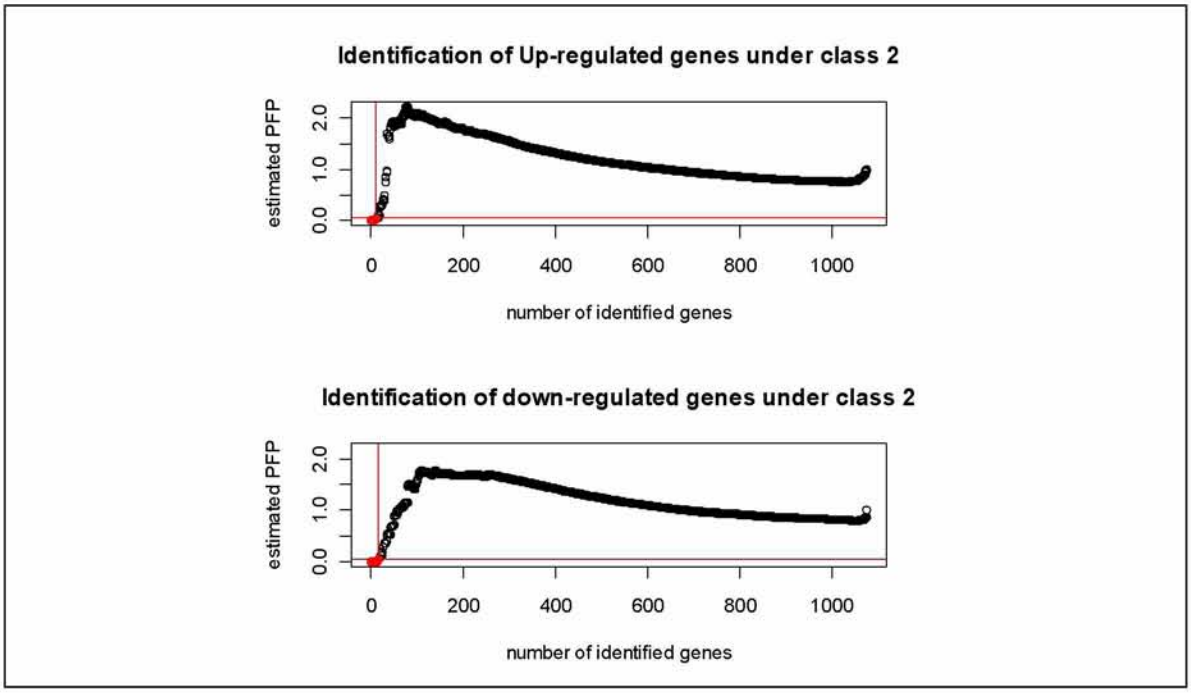


Εικόνα 3.3.8.3: Αποτελέσματα συνάρτησης topGene() για Σκλήρηνση κατά πλάκα ενάντια με Παχυσαρκία, με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια της Σκλήρησης κατά πλάκας ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Παχυσαρκίας.

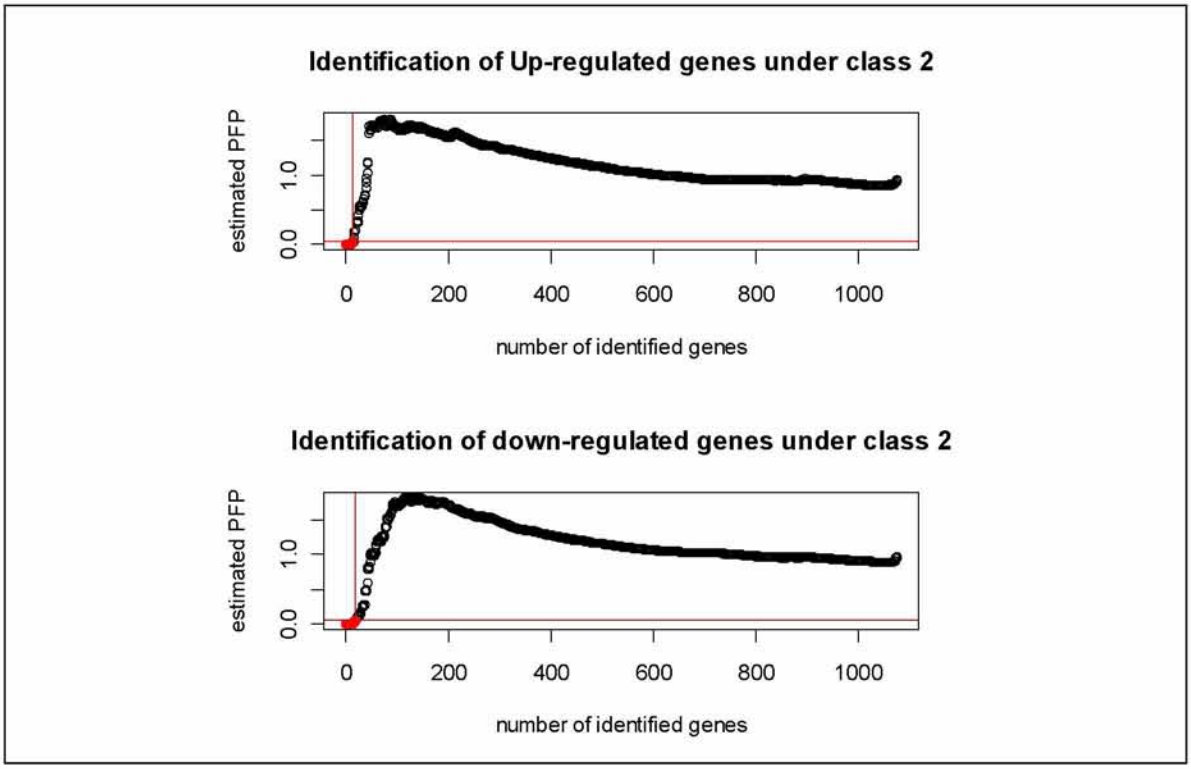
```
Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2

$Table1
      gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
PTPRC         31  39.1947                  0 0.0000   0e+00
MBP            2  39.2027                  0 0.0000   0e+00
CRYAB        388  66.8252                  0 0.0067   0e+00
PRKCA        462  78.8803                  0 0.0050   0e+00
APOE          16  84.2956                  0 0.0080   0e+00
IL4           18 104.8665                  0 0.0367   3e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
LEP           851  37.2412                Inf 0.0000   0e+00
UCP2          777  41.1114                Inf 0.0000   0e+00
MC4R          326  48.4108                Inf 0.0000   0e+00
PCSK1         367  55.7403                Inf 0.0000   0e+00
LPL           290  57.3904                Inf 0.0000   0e+00
PPARG         384  58.5313                Inf 0.0000   0e+00
MC3R          364  61.5680                Inf 0.0000   0e+00
POMC           74  62.4253                Inf 0.0000   0e+00
FTO           783  64.5430                Inf 0.0000   0e+00
ADRB2         216  64.6642                Inf 0.0000   0e+00
IRS2          611  75.5932                Inf 0.0009   0e+00
AGRP          128  76.0848                Inf 0.0008   0e+00
ADIPOQ        177  79.8774                Inf 0.0023   0e+00
RETN          452  87.6075                Inf 0.0043   1e-04
ADRB3         780  89.8093                Inf 0.0060   1e-04
NR3C1         240 106.9194                Inf 0.0269   5e-04
NPY           137 115.0453                Inf 0.0400   8e-04
```

3.3.9 Αποτελέσματα Υπέρτασης ενάντια σε Παχυσαρκία.

Εικόνα 3.3.9.1: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Υπέρταση ενάντια με Παχυσαρκία.



Εικόνα 3.3.9.2: Αποτελέσματα συνάρτησης plotRP(RP.out, cutoff = 0.05) για Υπέρταση ενάντια με Παχυσαρκία όπου RP.out=RPadvance.



Εικόνα 3.3.9.3: Αποτελέσματα συνάρτησης topGene() για Υπέρταση ενάντια με Παχυσαρκία , με ποσοστό ψευδώς θετικών προβλέψεων μικρότερο του 0.05. Στους δύο πίνακες παρουσιάζονται τα διαφορικά εκφρασμένα γονίδια. Στο πίνακα 1 δίνεται βαρύτητα στα γονίδια της Υπέρτασης ενώ στον πίνκακα 2 δίνεται βαρύτητα στα γονίδα της Παχυσαρκίας.

.

```
Table1: Genes called significant under class1 < class2

Table2: Genes called significant under class1 > class2

$Table1
        gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
BDKRB2         365  45.2114                  0 0.0000   0e+00
AGTR1          475  73.4329                  0 0.0000   0e+00
BMPR2            1  76.4460                  0 0.0000   0e+00
AGT            490  94.1016                  0 0.0125   0e+00
CYP11B2        919  95.3207                  0 0.0100   0e+00
EDN1             5  95.4345                  0 0.0083   0e+00
DRD1             8  97.1404                  0 0.0086   1e-04
NPPC            18 100.8481                  0 0.0100   1e-04
EDNRA            6 106.4570                  0 0.0122   1e-04
EDN2           172 107.2455                  0 0.0120   1e-04
NPPA             3 115.7886                  0 0.0209   2e-04
CAT             95 117.8940                  0 0.0217   2e-04
APOE           121 123.6823                  0 0.0338   4e-04
ADM             68 130.9646                  0 0.0400   5e-04

$Table2
       gene.index  RP/Rsum FC:(class1/class2)    pfp P.value
MC4R          434  26.9968                Inf 0.0000   0e+00
PCSK1         436  42.9819                Inf 0.0000   0e+00
MC3R          435  44.6174                Inf 0.0000   0e+00
LIPE          323  46.1992                Inf 0.0000   0e+00
IRS2          400  65.7690                Inf 0.0000   0e+00
LEP           217  66.8649                Inf 0.0000   0e+00
FTO           934  71.4914                Inf 0.0000   0e+00
LPL            72  79.5079                Inf 0.0000   0e+00
POMC           12  80.3085                Inf 0.0000   0e+00
SORBS1        858  83.7603                Inf 0.0000   0e+00
LEPR          776  89.7273                Inf 0.0018   0e+00
NR3C1          78  96.2836                Inf 0.0067   1e-04
ADRA2A        951 114.6890                Inf 0.0185   2e-04
PPARG         263 114.7863                Inf 0.0171   2e-04
AGRP           46 117.0213                Inf 0.0173   2e-04
NR1H3         388 122.2351                Inf 0.0256   4e-04
GHRL          169 130.7917                Inf 0.0329   5e-04
ADIPOQ        286 132.2387                Inf 0.0317   5e-04
UCP1         1056 136.1669                Inf 0.0332   6e-04
RETN          432 138.8115       5.378569e+17 0.0355   7e-04
```

ΚΕΦΑΛΑΙΟ 4: ΣΥΜΠΕΡΑΣΜΑΤΑ

4.1 Γενικά Συμπεράσματα.

Με την ανάπτυξη του κλάδου της Βιοπληροφορικής, έχουμε μία μεγάλη επανάσταση στον τρόπο με τον οποίο οι σύγχρονοι ερευνητές διεξάγουν τις διάφορες έρευνες τους για το βιολογικό/επιστημονικό πεδίο ενδιαφέροντος που έχουν. Η προσπάθεια όλων των επιστημών να δημιουργήσουν μια ηλεκτρονική εγκυκλοπαίδεια που να περιέχει μέσα όλες τις πληροφορίες που έχουν βρεθεί για το ανθρώπινο γονιδίωμα, έχει φέρει πολύ μεγαλύτερα αποτελέσματα απ' ότι περίμεναν. Τα διάφορα βιολογικά δεδομένα που τις απαρτίζουν αυξάνονται με απίστευτα ραγδαίους ρυθμούς. Η έλευση της Γονιδιωματικής έρευνας έχει προκαλέσει μεγάλης κλίμακας σύνολα δεδομένων και είναι προσανατολισμένη προς την ανακάλυψη των βιολογικών σημάτων και την κατανόηση των οδών και των ενώσεων. Συνήθως αυτά τα σύνολα περιλαμβάνουν πληροφορίες σε εκατοντάδες – χιλιάδες βιολογικές μεταβλητές σε μια σειρά από δείγματα.

Μέσα από τα διάφορα στάδια της ανάπτυξης όλης αυτής της μεγάλης ιδέας λόγω του μεγάλου όγκου δεδομένων που συγκεντρώνεται στις διάφορες ηλεκτρονικές βάσεις δεδομένων, υπήρξε η ανάγκη για δημιουργία αυτόματων μηχανών αναζήτησης, ούτως ώστε ο κάθε ερευνητής να μπορεί να βρίσκει εύκολα και γρήγορα αυτό που αναζητά με το πάτημα ενός κουμπιού. Αφού δημιουργήθηκε και εδραιώθηκε μία μεγάλη γκάμα από αυτόματες μηχανές αναζήτησης, οι οποίες ανακτούν δεδομένα μέσα από τις υπάρχουσες βάσεις δεδομένων, τότε εμφανίστηκε μία καινούργια ιδεολογία μηχανών αναζήτησης. Οι εν λόγω μηχανές πέρα από την απλή αναζήτηση που παρέχουν, ανέπτυξαν την ικανότητα μέσα από τις διάφορες μεθοδολογίες και τεχνικές που χρησιμοποιούν (Διανυσματικές μηχανές στήριξης (SVM), Νευρωνικά δίκτυα, Δένδρα απόφασης, Δενδρικές αναζητήσεις, Μπεϊσιανά δίκτυα, Έλεγχος Man Whitney U, t-test, Δίκτυα συσχετισμού γονιδίων (Y2H,KnowledgeNet), Μετασχηματισμοί z-score) να λειτουργούν πιο έξυπνα, συνδυάζοντας τα διάφορα δεδομένα που αναζητούν μεταξύ τους και παρέχοντας στους χρήστες μια πιο ολοκληρωμένη εικόνα γύρο από το αντικείμενο που τους ενδιαφέρει.

Χαρακτηριστικά παραδείγματα περιλαμβάνουν μεθόδους ανάλυσης των μικροσυστοιχιών, που βασίζονται στη σκιαγράφηση της έκφρασης των γονιδίων. Η φασματοσκοπία μάζας βασίζεται στην πρωτεωμική και στο ευρύ γονιδίωμα με τη βοήθεια γενετικών μελετών σύνδεσης. Τα αποτελέσματα αυτών των έξυπνων μηχανών αναζήτησης και ιεράρχησης, παρεμποδίζονται όχι μόνο από την πολυπλοκότητα των βιολογικών προβλημάτων αλλά και από τα μικρά μεγέθη των δειγμάτων και τα

αντιφατικά αποτελέσματα μεταξύ των μελετών. Συχνά οι ενώσεις για κάθε μία βιολογική μεταβλητή τείνουν να είναι μάλλον αδύναμες.

Η κατάσταση αυτή δημιουργεί την ανάγκη συνδυασμού των αποτελεσμάτων από όλους τους διαφορετικούς ιεραρχικούς αλγόριθμους αναζήτησης που υπάρχουν, για να προσπαθήσουν να μεγιστοποιήσουν την ανίχνευση των πιο σωστών αποτελεσμάτων. Ο συνδυασμός των δεδομένων είναι ωστόσο μια περαιτέρω πρόκληση, δεδομένου ότι τα διαθέσιμα σύνολα δεδομένων μπορούν να έχουν ληφθεί με διαφορετικό τρόπο πειραματικών συνθηκών, διαφορετικές τεχνικές ανάλυσης, ή ακόμα και διαφορετικές βιολογικές παράμετρους (π.χ. Δημοσιευμένη Βιβλιογραφία, Μοριακό & Λειτουργικό επίπεδο πρωτεϊνών, Κωδικοποιημένη ακολουθία πρωτεϊνών, Οντολογίες γονιδίων, Πρότυπα έκφρασης (Μικροσυστοιχίες), Νουκλεοτιδικοί πολυμορφισμοί, Πρωτεϊνικές περιοχές, Φυλογενετικότητα, Παράλογα, Ομόλογα, Φαινοτυπικές παραλλαγές, Πειραματικά δεδομένα από ζώα, Λειτουργικός σχολιασμός γονιδίων, Χαρακτηριστικά ακολουθίας γονιδίου, Αριθμός εξονίων, Δίκτυα αλληλεπίδρασης πρωτεϊνών (PPIs), Ανάλυση μικροσυστοιχιών). Οι υπό μελέτη βιολογικές μεταβλητές μπορεί επίσης να επικαλύπτονται, αλλά δεν πρέπει να ταυτίζονται με όλες τις μελέτες .

Στην παρούσα εργασία προτείνουμε ένα επόμενο βήμα στην εξέλιξη της όλης ιδέας της Βιοπληροφορικής κοινότητας. Παρουσιάζουμε ένα τρόπο συνδυασμού των αποτελεσμάτων των αυτόματων έξυπνων ιεραρχικών μηχανών αναζήτησης και επικεντρωνόμαστε στα υποψήφια γονίδια που φαίνεται να σχετίζονται με ασθένειες. Για τον λόγο αυτό χρησιμοποιούμε τους αλγόριθμους ιεραρχικής αναζήτησης Prospectr, Suspects, Gene-Prospectr, Phenopred, Posmed, Candid, SNPs3d, FITsSNPs και GeneCards, οι οποίοι μας επιστρέφουν τα υποψήφια γονίδια απλά εισάγοντας τις λέξεις κλειδιά που σχετίζονται με την ασθένεια που αναζητούμε. Οι διάφορες ασθένειες για τις οποίες διεξάχθηκε η εργασία μας είναι ο Καρκίνος του μαστού, Διαβήτης τύπου Ι, Διαβήτης τύπου ΙΙ, Υπέρταση, Σκλήρυνση κατά πλάκα και Παχυσαρκία.

Αρχικά παράγουμε τις λίστες με τα υποψήφια γονίδια που σχετίζονται με τις διάφορες ασθένειες, κάνοντας χρήση της ιεραρχικής ταξινόμησης που δέχεται κάθε γονίδιο μέσα από κάθε αλγόριθμο. Είναι πολύ αξιόπιστες, αφού παράγονται από εννέα διαφορετικούς ιεραρχικούς αλγόριθμους αναζήτησης, που και αυτοί με τη σειρά τους εφαρμόζουν πολλές τεχνικές/ μεθόδους που έχουν υλοποιηθεί μέχρι σήμερα με επιτυχία, αλλά και κάνουν χρήση πληροφοριών από όλες τις διαθέσιμες βάσεις δεδομένων που

υπάρχουν. Πέρα από τη δημιουργία των λιστών με όλα τα υποψήφια γονίδια, τα δεδομένα μας επεξεργάζονται με δύο στατιστικές μεθόδους ανάλυσης το RankProd (Κεφάλαιο 2.6) και Metradisc (Κεφάλαιο 2.7), οι οποίες αναλύουν τα δεδομένα μας και μας παρουσιάζουν τα πιο υψηλά υποψήφια γονίδια που σχετίζονται με τις ασθένειες που αναζητούμε.

Μέσα από την πορεία της επιστημονικής κοινότητας δεν είναι η πρώτη φορά που εφαρμόζεται η ιδέα της εκτέλεσης πολυσύνθετων αναζητήσεων, και συγκέντρωση αποτελεσμάτων με βάση πολλών διαφορετικών βιολογικών παραμέτρων. Παρόμοια εργασία έχει εκτελεστεί στον παρελθόν με την μόνη διαφορά ότι τα δεδομένα της συγκεκριμένης εργασίας προέρχονταν από τρείς αλγόριθμους αναζήτησης [127].

Γενικότερα σαν ιδέα η διαδικασία που εκτελέσαμε είναι πολύ κοντινή με την μέθοδο μετα-ανάλυσης γονιδίων[128]. Στην μετα-ανάλυση με παρόμοιο τρόπο, για να δούμε κατά πόσο πολύ ένα γονίδιο συνδέεται με μία ασθένεια αναλύουμε τα πρότυπα έκφρασης πολλών τύπων ιστών ούτως ώστε να δούμε κατά πόσο πολύ το γονίδιο σχετίζεται ή δεν σχετίζεται με την νόσο.

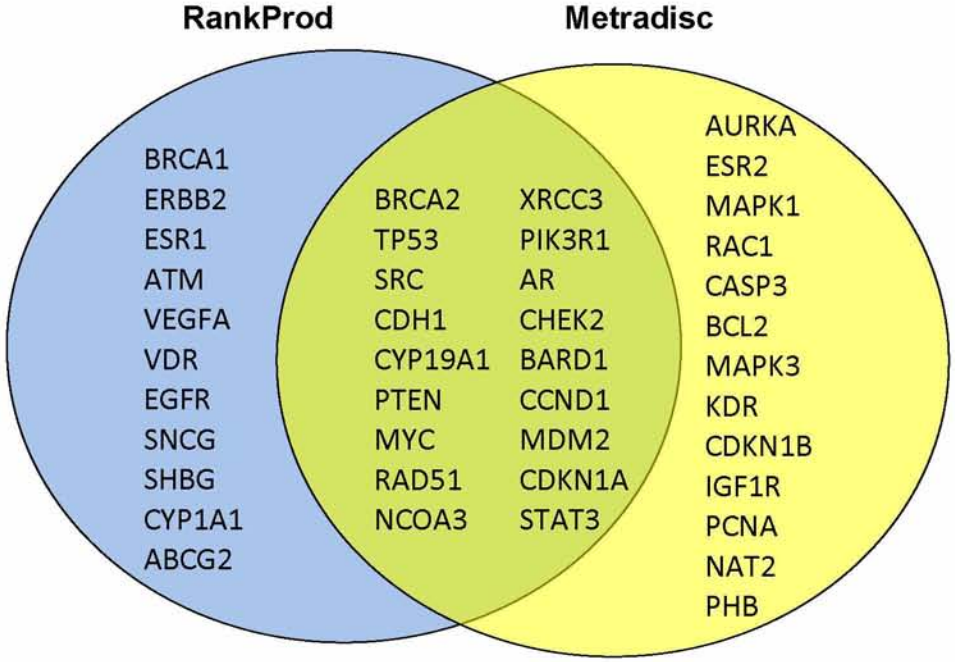
4.2 Συμπεράσματα με βάση τα αποτελέσματα

4.2.1. Υψηλά υποψήφια γονίδια «Top Genes» που σχετίζονται με καρκίνο του μαστού.

Στην Εικόνα 4.2.1.1 και στον πίνακα 4.2.1.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια καρκίνο του μαστού.

Με βάση τα αποτελέσματα που εξάγουμε για τον καρκίνο του μαστού, κάποια από αυτά τα γονίδια είναι πλέον ευρέως διαδεδομένα ότι συνδέονται με τον καρκίνο του μαστού λόγω του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί γι' αυτά, όπως το BRCA2[129, 130], CYP19A1[131, 132], PTEN[129, 130] ,RAD51[133, 134], CHEK2 [135, 136], CCND1[137, 138], MDM2[139] και SRS[140, 141]. Μερικά από αυτά έχουν συσχετιστεί και πειραματικά.

Εικόνα 4.2.1.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Καρκίνο του μαστού.



Κάποια άλλα από αυτά τα γονίδια, πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με τον Καρκίνο του μαστού, για παράδειγμα τα γονίδια CDH1[142, 143], MYC[144, 145], XRCC3[146, 147], AR[148, 149], BARD1[150, 151], STAT3[152, 153] και TP53[154, 155] όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους

τρόπους η σύνδεση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια στο προσεχές μέλλον.

Σε κάποια άλλα γονίδια όπως NCOA3[156, 157], PIK3R1 και CDKN1A που φαίνεται να σχετίζονται με τον καρκίνο του μαστού, η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα ή στις περιπτώσεις που έχουν γίνει έρευνες είναι λίγες και αντιφατικές. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια του Καρκίνου του μαστού, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Γενικότερα είχαμε πολύ καλά αποτελέσματα, με γονίδια τα οποία, αναμένονταν να ανιχνευτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.1.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Καρκίνο του μαστού και οι βιολογικές τους παράμετροι.

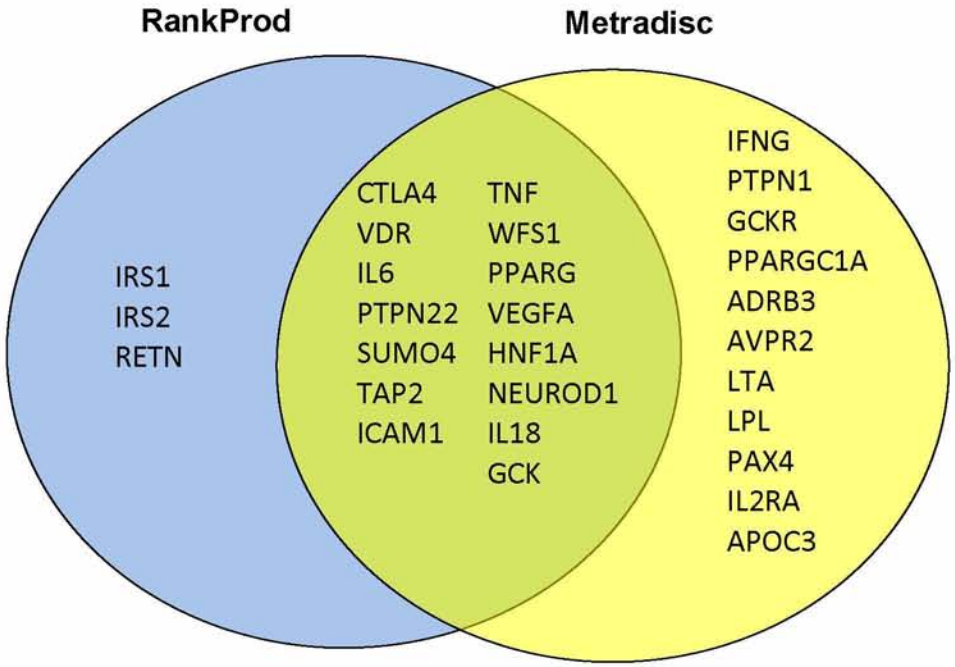
Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Πρότυπα έκφρασης	Αναφορές PMID
BRCA2	2783	6	13q12.3	9757	8640236
TP53	3308	11	17p13.1	29464	20708154
SRC	22910	395	20q12-q13	139914	20460377
CDH1	1347	3	16q22.1	6706	20716965
CYP19A1	365	1	15q21.1	4262	17119036
PTEN	5896	17	10q23.3	179853	20637195
MYC	34753	36	8q24.21	237335	20551172
RAD51	2884	11	15q15.1	11171	20461453
NCOA3	157	1	20q12	5293	20422428
XRCC3	466	1	14q32.3	1824	20549576
PIK3R1	231	1	5q13.1	5018	20530665
AR	94564	405	Xq12	296320	20534771
CHEK2	326	1	22q12.1	1958	20496165
BARD1	419	3	2q34-q35	2040	20453000
CCND1	1572	1	11q13	4919	20701069
MDM2	4501	15	12q14.3-q15	7879	20582981
CDKN1A	1847	9	6p21.2	10413	20524403
STAT3	4584	1	17q21.31	87448	16081048

4.2.2. Υψηλά υποψήφια γονίδια «Top Genes» που σχετίζονται με Διαβήτη τύπου Ι.

Στην Εικόνα 4.2.2.1 και στον πίνακα 4.2.2.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια Διαβήτη τύπου Ι.

Με βάση τα αποτελέσματα που εξάγουμε για τον Διαβήτη τύπου Ι, κάποια γονίδια είναι ευρέως διαδεδομένα ότι σχετίζονται με αυτή την ασθένεια λόγο του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί για αυτά, όπως το IL6[158, 159], VDR[160, 161], TNF[162, 163] και VEGFA[164]. Κάποια άλλα γονίδια πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με τον Διαβήτη τύπου Ι, για παράδειγμα τα γονίδια CTLA4[165] και ICAM1[166] όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους τρόπους η συσχέτιση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια στο προσεχές μέλλον .

Εικόνα 4.2.2.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Διαβήτη τύπου Ι.



Σε κάποια άλλα γονίδια όπως PTPN22, TAP2, WFS1[167], PARG[168],HNF1A, NEUROD1[169, 170], IL18 και GCK που παρουσιάζονται να σχετίζονται με τον Διαβήτη τύπου ΙΙ στην εργασία μας, αλλά η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα. Σε παρόμοιο στάδιο έρευνας παρουσιάζεται το γονίδιο SUMO4, για το οποίο δεν έχουν εκτελεστεί αρκετές έρευνες και μέσα από αυτές τις λίγες που βρέθηκαν, έχουμε αντιφατικά αποτελέσματα. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε περισσότερη έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια του Διαβήτη τύπου Ι, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Για τα γονίδια IL18 και GCK που όπως έχουμε προαναφέρει πιο πάνω, μέχρι σήμερα δεν έχει αποδειχθεί να σχετίζονται με το διαβήτη τύπου Ι και αποτελούν πτυχιακές εργασίες φοιτητών του Πανεπιστημίου Στερεάς Ελλάδος όπου προσπαθούν να αποδείξουν κατά πόσο σχετίζονται ή όχι με την ασθένεια αυτή. Σχετικά με το γονίδιο IL18 έχει αποδειχθεί η συσχέτιση του με τον διαβήτη τύπου Ι και παρουσιάστηκε σε συνέδριο που πραγματοποιήθηκε πρόσφατα. Ευελπιστούμε να έχουμε θετικά αποτελέσματα στην σύνδεση και του GCK με τον Διαβήτη τύπου Ι και να επιβεβαιώνει και αυτό με την σειρά του την παρούσα εργασία. Γενικά μέσα από όλη την διαδικασία που εκτελέστηκε είχαμε πολύ καλά αποτελέσματα, με γονίδια τα οποία αναμένονταν να ανιχνευτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.2.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Διαβήτη τύπου Ι και οι βιολογικές τους παράμετροι.

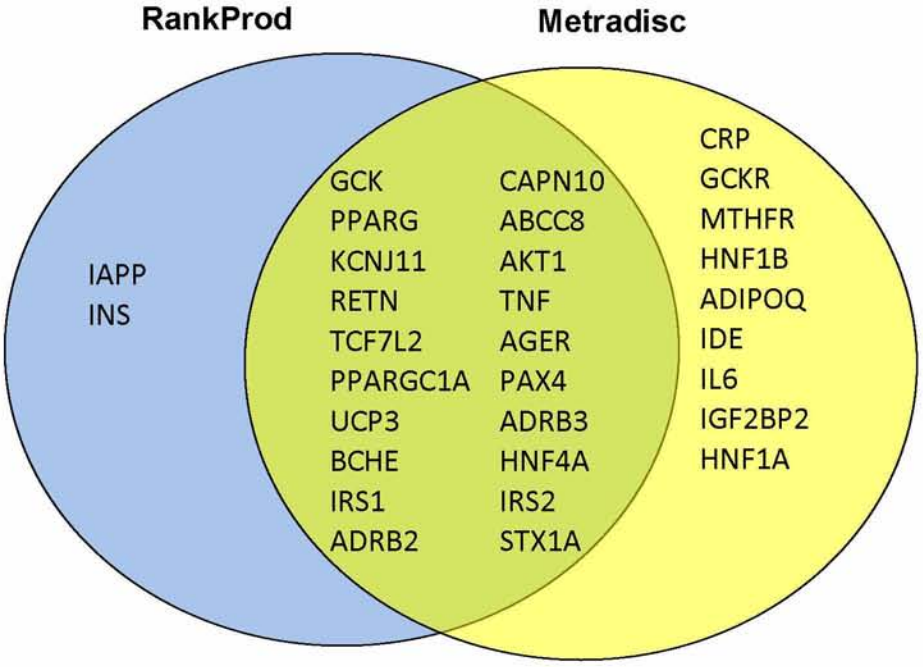
Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Προφίλ έκφρασης	Αναφορές PMID
CTLA4	1542	1	2q33	2920	18056379
VDR	1522	2	12q13.11	83762	12843155
IL6	3243	5	7p21	14893	-
PTPN22	497	1	1p13.2	3702	17054449
SUMO4	66	1	6q25	1389	17554341
TAP2	588	3	6p21.3	4997	17192492
ICAM1	2037	1	19p13.3-p13.2	3262	9566857
TNF	45099	71	6p21.3	371473	19477545
WFS1	45099	71	4p16	371473	18060660
PPARG	1047	1	3p25	59785	18091023
VEGFA	812	1	6p12	5171	11340407
HNF1A	260	1	12q24.2	25045	19388975
NEUROD1	404	1	2q32	47951	16357810
IL18	645	1	2q32	24824	10415018
GCK	483	10	7p15.3-p15.1	5183	14517946

4.2.3. Υψηλά υποψήφια γονίδια «Top Genes» που σχετίζονται με Διαβήτη τύπου ΙΙ.

Στην Εικόνα 4.2.3.1 και στον πίνακα 4.2.3.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια Διαβήτη τύπου ΙΙ.

Με βάση τα αποτελέσματα που εξάγουμε για τον Διαβήτη τύπου ΙΙ, κάποια γονίδια είναι πλέον, ευρέως διαδεδομένα ότι σχετίζονται με αυτή την ασθένεια λόγο του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί για αυτά, όπως το GCK[171, 172], PPARG[173, 174], KCNJ11[175, 176], PPARGC1A[177], IRS1[178, 179], CAPN10[180, 181], ABCC8[176, 182], TNF[183, 184], HNF4A[185] και TCF7L2[186, 187]. Πολλά από αυτά τα γονίδια η σύνδεση τους με την εν λόγο ασθένεια έχει αποδειχθεί πειραματικά.

Εικόνα 4.2.3.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Διαβήτη τύπου ΙΙ.



Κάποια άλλα γονίδια πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με τον Διαβήτη τύπου Ι, όπως τα γονίδια UCP3[188, 189], AGER[190], PAX4[191, 192], ADRB3[193, 194], IRS2[195] και ADRB2[194] όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους τρόπους η συσχέτιση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια στο προσεχές μέλλον. Σε κάποια άλλα γονίδια όπως RETN, BCHE και AKT1, STX1A που παρουσιάζονται να σχετίζονται με τον Διαβήτη τύπου ΙΙ στην εργασία μας, η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε περισσότερη έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια του Διαβήτη τύπου Ι, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Γενικότερα είχαμε πολύ καλά αποτελέσματα, μέσα από τα οποία κάποια γονίδια αναμένονταν να ανιχνευτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.3.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με τον Διαβήτη τύπου ΙΙ και οι βιολογικές τους παράμετροι.

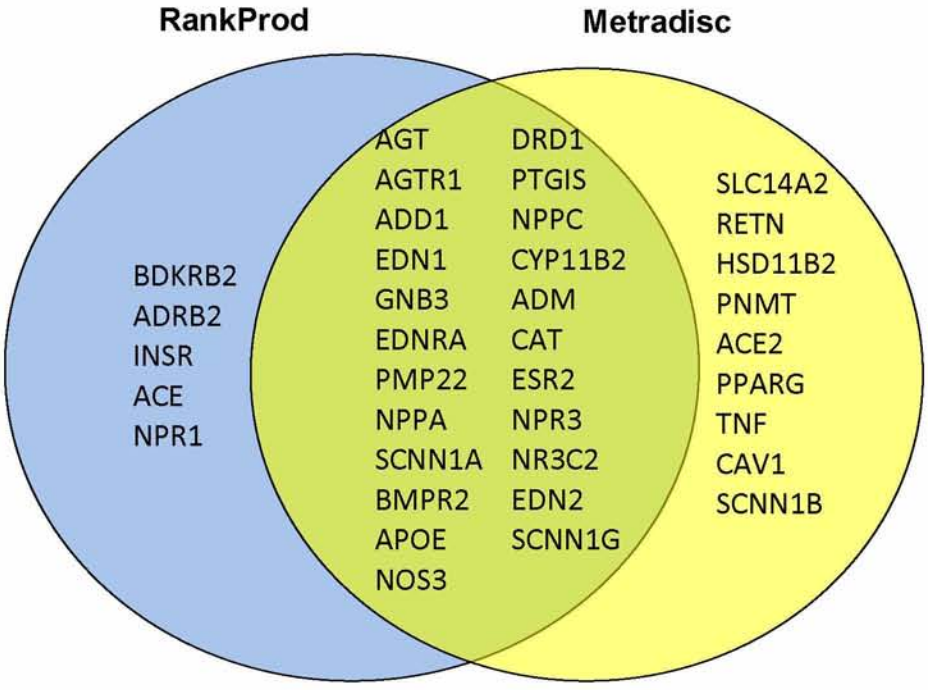
Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Προφίλ έκφρασης	Αναφορές PMID
GCK	483	10	7p15.3-p15.1	5183	14517946
PPARG	1047	1	3p25	59785	18091023
KCNJ11	319	1	11p15.1	2165	15448106
RETN	139	1	19p13.2	1608	20346233
TCF7L2	500	2	10q25.3	6823	17563454
PPARGC1A	269	1	4p15.1	3807	16435105
UCP3	433	2	11q13	3507	16644712
BCHE	270	1	3q26.1-q26.2	2731	11793025
IRS1	1207	2	2q36	3260	19557384
ADRB2	655	1	5q31-q32	2171	18249219
CAPN10	124	1	2q37.3	3245	16857402
ABCC8	177	1	11p15.1	3406	17259403
AKT1	2482	5	14q32.32	51534	11508278
TNF	45099	71	6p21.3	371473	16260352
AGER	934	1	6p21.3	2936	18796298
PAX4	265	1	7q32	2733	11723072
ADRB3	93	1	8p12	28739	16444766
HNF4A	278	1	20q13.12	73814	15793260
IRS2	500	1	13q34	2832	9495343
STX1A	175	2	7q11.23	2660	17912268

4.2.4. Υψηλά υποψήφια γονίδια που «Top Genes» σχετίζονται με Υπέρταση.

Στην Εικόνα 4.2.4.1 και στον πίνακα 4.2.4.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια Υπέρταση.

Με βάση τα αποτελέσματα που εξάγουμε για την Υπέρταση, τα γονίδια AGT[196, 197], BMPR2[198, 199], CYP11B2[200] και ADD1[201], GNB3[202].είναι πλέον, ευρέως διαδεδομένα ότι σχετίζονται με αυτή την ασθένεια λόγο του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί. Αξιοσημείωτο είναι ότι το γονίδιο GNB3[202] και η σχετική έρευνα σύνδεσης του με την ασθένεια της Υπέρτασης έχει πραγματοποιηθεί από τον επίκουρο καθηγητή του Πανεπιστημίου Στερεάς Ελλάδος κ. Μπάγκο Παντελή, ο οποίος είναι επιβλέπων καθηγητής και της παρούσας εργασίας.

Εικόνα 4.2.4.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Υπέρταση.



Κάποια άλλα γονίδια όπως AGTR1[203, 204], NPPA[205], EDNRA[206], APOE[207, 208], NOS3[209, 210], DRD1[211], ADM[212], CAT[213, 214], SCNN1G[215] και EDN1[216] πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με την Υπέρταση, όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους τρόπους η

συσχέτιση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια . Σε κάποια άλλα γονίδια όπως PMP22, PTGIS, NPPC, ESR2, NPR3, NR3C2, EDN2 και SCNN1A που παρουσιάζονται να σχετίζονται με την Υπέρταση στην εργασία μας, η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε περισσότερη έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια της Υπέρτασης, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Γενικά μέσα από όλη την διαδικασία που εκτελέστηκε είχαμε πολύ καλά αποτελέσματα, με γονίδια τα οποία αναμένονταν να ανιχνευτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.4.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Υπέρταση και οι βιολογικές τους παράμετροι.

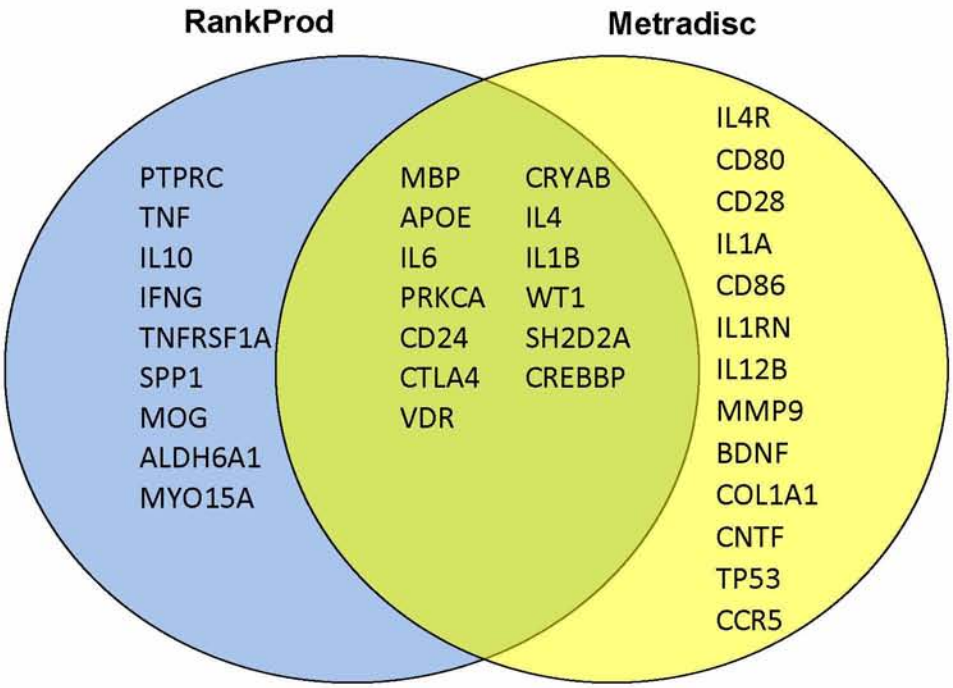
Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Προφίλ έκφρασης	Αναφορές PMID
AGT	15470	7	1q42-q43	63088	9049480
AGTR1	159	3	3q24	3542	15332573
ADD1	266	2	4p16.3	5911	17003363
EDN1	209	1	6p24.1	2640	18288492
GNB3	145	1	12p13	2442	16487269
EDNRA	98	2	4q31.22	4944	14616768
PMP22	465	21	17p12	16968	8554058
NPPA	172	1	1p36.21	2281	19219041
SCNN1A	82	1	12p13	3035	12530930
BMPR2	236	1	2q33-q34	49588	12358323
APOE	4517	1	19q13.2	139202	18297189
NOS3	953	2	7q36	3290	17762636
DRD1	210	4	5q35.1	2799	10948075
PTGIS	62	1	20q13.13	3679	19265782
NPPC	93	1	2q24-qter	1884	12452325
CYP11B2	138	1	8q21-q22	2065	9931115
ADM	3512	3	11p15.4	2391	16212982
CAT	86213	43	11p13	16477	15735318
ESR2	228	1	14q23.2	5002	15894829
NPR3	61	2	5p14-p13	6763	12872042
NR3C2	164	1	4q31.1	1637	19325532
EDN2	51	1	1p34	1955	12884521
SCNN1G	40	1	16p12	2081	17698725

4.2.5. Υψηλά υποψήφια γονίδια που σχετίζονται «Top Genes» με Σκλήρυνση κατά πλάκας.

Στην Εικόνα 4.2.5.1 και στον πίνακα 4.2.5.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια Σκλήρυνση κατά πλάκα.

Με βάση τα αποτελέσματα που εξάγουμε για την Σκλήρυνση κατά πλάκα κάποια γονίδια είναι πλέον, ευρέως διαδεδομένα ότι σχετίζονται με αυτή την ασθένεια λόγο του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί για αυτά, όπως το MBP[217, 218], CRYAB[219] και APOE[220]. Από την άλλη μεριά κάποια άλλα γονίδια πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με την Σκλήρυνση κατά πλάκα, για παράδειγμα τα γονίδια IL6[221], CD24[222, 223], VDR[224], IL4[225, 226], SH2D2A[227, 228] και PRKCA[229, 230] όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους τρόπους η συσχέτιση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια στο προσεχές μέλλον

Εικόνα 4.2.5.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Σκλήρυνση κατά πλάκας.



Επίσης σε κάποια άλλα γονίδια όπως το IL1B, CREBBP και WT1 που παρουσιάζονται να σχετίζονται με την Σκλήρυνση κατά πλάκα στην εργασία μας, η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα. Σε παρόμοιο στάδιο έρευνας παρουσιάζεται το γονίδιο CTLA4[231, 232], για το οποίο δεν έχουν εκτελεστεί αρκετές έρευνες και μέσα από αυτές τις λίγες που βρέθηκαν, έχουμε αντιφατικά αποτελέσματα. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε περισσότερη έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια της Σκλήρυνσης κατά πλάκα, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Γενικότερα είχαμε πολύ καλά αποτελέσματα, με γονίδια τα οποία αναμένονταν να ανιχνευτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.5.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Σκλήρυνση κατά πλάκας και οι βιολογικές τους παράμετροι.

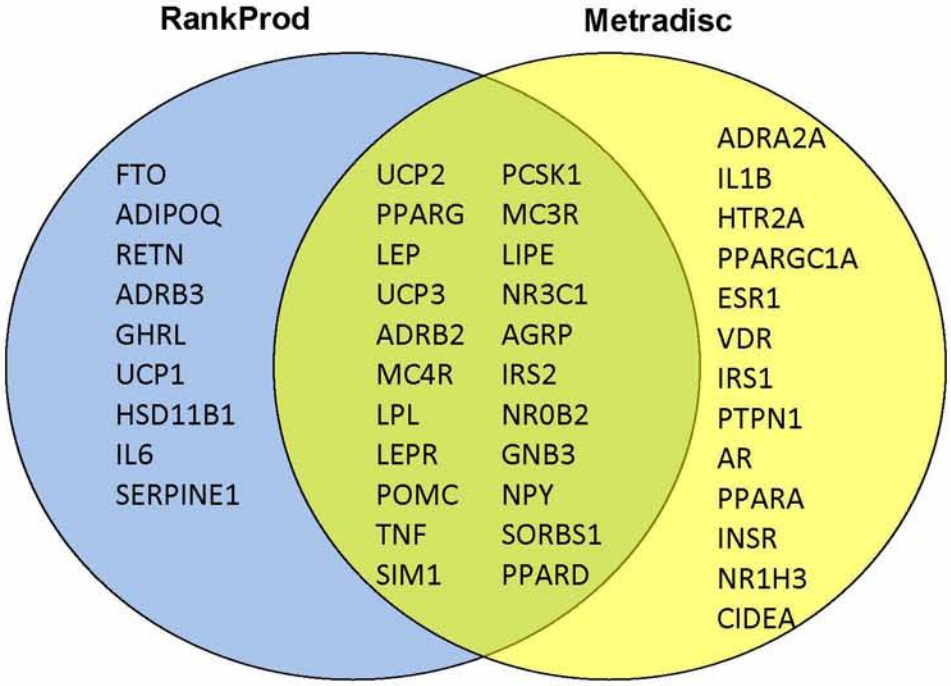
Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Προφίλ έκφρασης	Αναφορές PMID
MBP	40	1	18q23	2081	12618862
APOE	4517	1	19q13.2	139202	19140315
IL6	3244	5	7p21	14893	11196678
PRKCA	370	2	17q22-q23.2	6377	15155525
CD24	1408	2	6q21	50950	16631259
CTLA4	1544	1	2q33	2920	10408973
VDR	1522	2	12q13.11	83762	19894309
CRYAB	225	1	11q22.3-q23.1	3462	17010329
IL4	24555	2	5q31.1	195031	18239607
IL1B	654	1	2q14	102784	14664464
WT1	1902	6	11p13	491197	-
SH2D2A	30	1	1q21	1803	11528519
CREBBP	379	5	16p13.3	10319	18776589

4.2.6. Υψηλά υποψήφια γονίδια «Top Genes» που σχετίζονται με Παχυσαρκία.

Στην Εικόνα 4.2.6.1 και στον πίνακα 4.2.6.2 που ακολουθούν παρουσιάζονται και περιγράφονται αναλυτικά τα υψηλά υποψήφια γονίδια που σχετίζονται με την ασθένεια Παχυσαρκία.

Με βάση τα αποτελέσματα που εξάγουμε για την Παχυσαρκία, κάποια γονίδια όπως το UCP2[233, 234] και MC4R[235, 236] είναι πλέον, ευρέως διαδεδομένα ότι σχετίζονται με αυτή την ασθένεια λόγο του μεγάλου αριθμού ερευνών που έχουν δημοσιευτεί για αυτά. Από την άλλη μεριά κάποια άλλα γονίδια πρόσφατα έχουν χαρακτηριστεί να σχετίζονται με την Παχυσαρκία, για παράδειγμα τα γονίδια PPARG[237], UCP3[238, 239], LPL[240], POMC[241, 242], TNF[243, 244], SIM1[245, 246], PCSK1[247, 248], AGRP[249, 250], IRS2[251, 252], GNB3[253], NPY[249, 250], SORBS1[254] και LEP[255, 256] όπου με βάση τις δημοσιευμένες έρευνες έχει βρεθεί με διάφορους τρόπους η συσχέτιση τους με την ασθένεια, και έτσι αναμένουμε περισσότερη έρευνα και έμφαση σε αυτά τα γονίδια στο προσεχές μέλλον .

Εικόνα 4.2.6.1: Συνδυασμός αποτελεσμάτων RankProd και Metradisc για τα αποτελέσματα με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Παχυσαρκία.



Υπάρχουν και περιπτώσεις γονιδίων όπως το MC3R, NR3C1, NR0B2, PPARD και LIPE που παρουσιάζονται να σχετίζονται με την Παχυσαρκία στην εργασία μας, αλλά η επιστημονική κοινότητα ακόμη δεν έχει εκτελέσει καμία έρευνα. Σε παρόμοιο στάδιο έρευνας παρουσιάζεται τα γονίδια ADRB2[252, 257] και LEPR[258, 259] για τα οποία δεν έχουν εκτελεστεί αρκετές έρευνες και μέσα από αυτές τις λίγες που βρέθηκαν, έχουμε αντιφατικά αποτελέσματα. Άρα σε αυτές τις περιπτώσεις γονίδιων χρειαζόμαστε περισσότερη έρευνα για να μπορέσουμε να κατανοήσουμε μέσα από ποιους γενετικούς παράγοντες εκδηλώνεται η ασθένεια της Παχυσαρκίας, αλλά και τους τρόπους μέσα από τους οποίους μπορεί να αλληλεπιδρούν κάποια γονίδια μεταξύ τους για την εκδήλωση της. Γενικά μέσα από όλη την διαδικασία που εκτελέστηκε είχαμε πολύ καλά αποτελέσματα, με γονίδια τα οποία αναμένονταν να προβλεφτούν αλλά και προτείνουμε κάποια άλλα τα οποία θα πρέπει να ερευνηθούν στα επόμενα χρόνια.

Πίνακας 4.2.6.2 : Αναλυτικός πίνακας με τα υψηλά υποψήφια γονίδια που σχετίζονται με την Παχυσαρκία και οι βιολογικές τους παράμετροι.

Γονίδιο	Δημοσιεύσεις PubMed	Ομόλογα γονίδια	Περιοχή (locus)	Προφίλ έκφρασης	Αναφορές PMID
UCP2	549	2	11q13	4133	19769793
PPARG	1047	1	3p25	59785	8647948
LEP	1603	5	7q31.3	25256	20520958
UCP3	433	2	11q13	3507	16006788
ADRB2	655	1	5q31-q32	2171	16493638
MC4R	400	1	18q22	1035	18697794
LPL	2337	4	8p22	4699	20429872
LEPR	1013	2	1p31	65498	19142102
POMC	3285	1	2p23.3	2491	11151766
TNF	45099	71	6p21.3	371473	16741264
SIM1	210	2	6q16.3-q21	2693	16741264
PCSK1	92	1	5q15-q21	2152	19528091
MC3R	155	1	20q13.2-q13.3	1620	10702772
LIPE	233	1	19q13.2	2985	7489032
NR3C1	234	1	5q31.3	6524	16741264
AGRP	581	1	16q22	1676	16568137
IRS2	500	1	13q34	2832	12687350
NR0B2	90	1	1p36.1	2362	11136233
GNB3	145	1	12p13	2442	16487269
NPY	2326	6	7p15.1	1983	20531438
SORBS1	74	1	10q23.33	6014	12849814
PPARD	111	1	6p21.2	4107	8647948

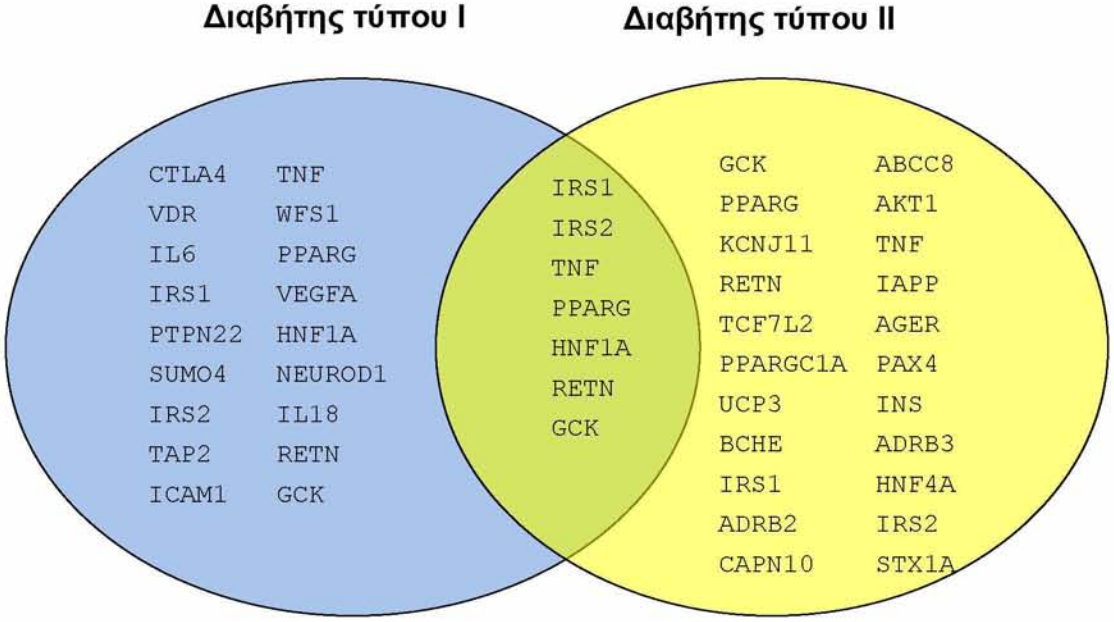
4.3 Συμπεράσματα - Ασθένειες Πολυπαραγοντικής Αιτιολογίας.

Οι ασθένειες πολυπαραγοντικής αιτιολογίας [260, 261] αποτελούν ένα νέο σύνολο στον τομέα της γενετικής επιδημιολογίας μέσα από τον οποίο οι ερευνητές προσπαθούν να συνδέσουν τις ασθένειες οι οποίες συσχετίζονται μεταξύ τους και που μπορεί να λειτουργούν ως μία ασθένεια στο σύνολο, η οποία να εκφράζεται διαφορετικά ανάλογα από κάποιους γενετικούς και περιβαλλοντικούς παράγοντες.

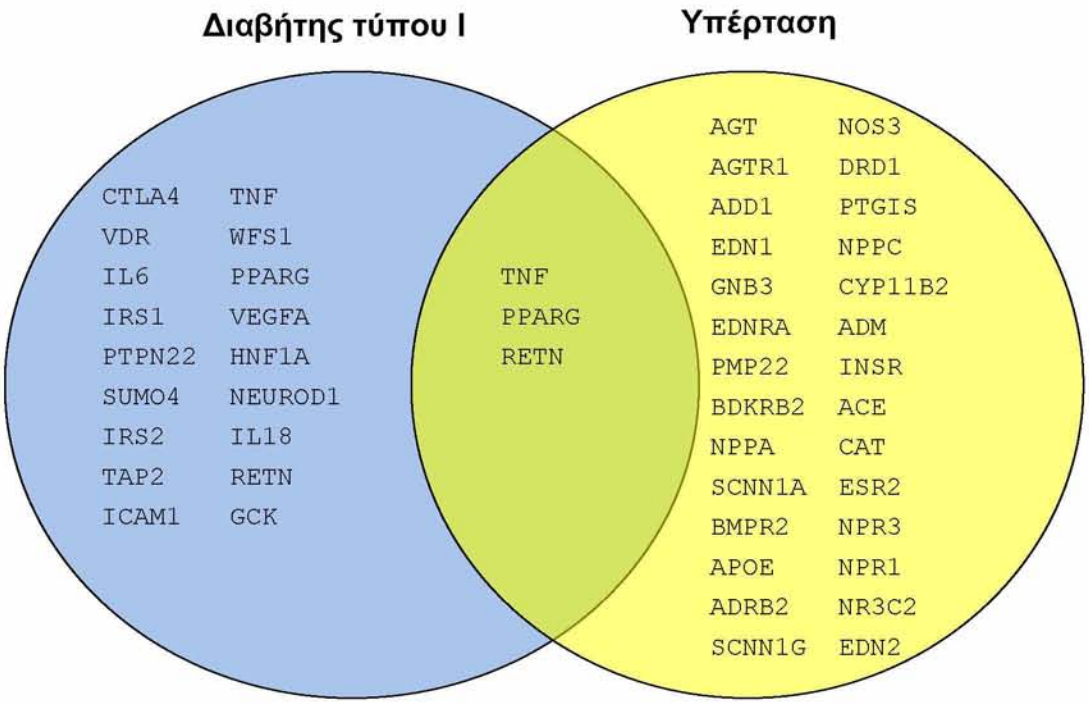
Μετά από σύγκριση των υψηλά υποψήφιων γονιδίων που βρήκαμε από τις μεθόδους στατιστικής ανάλυσης RankProd (Κεφάλαια 3.1-3.2) και Metradisc (Κεφάλαια 3.1-3.2), ανακαλύψαμε ότι κάποιες ασθένειες μοιράζονταν συγκεκριμένα γονίδια μεταξύ τους. Με βάση τους συνδυασμούς ανά δύο ασθένειες που πήραμε, βρίσκαμε ασθένειες που σχετίζονταν περισσότερα ή ασθένειες που σχετίζονταν λιγότερο κάθε φορά, ανάλογα με τα κοινά υψηλά υποψήφια γονίδια που παρουσίαζαν.

Στις επόμενες σελίδες που ακολουθούν παρουσιάζουμε τους συνδυασμούς ασθενειών που επιλέξαμε σε σχήματα και τα κοινά / μη κοινά γονίδια που παρουσιάζουν μεταξύ τους. Τα γονίδια που παρουσιάζονται στα σχήματα προέρχονται από την μέθοδο στατιστικής ανάλυσης RankProd. Για σκοπούς εξοικονόμησης χώρου δεν παρουσιάζονται όλα τα υψηλά υποψήφια γονίδια της μεθόδου Metradisc στα σχήματα παρά μόνο τα κοινά υψηλά υποψήφια γονίδια που βρέθηκε να είναι από κοινού μεταξύ των ασθενειών.

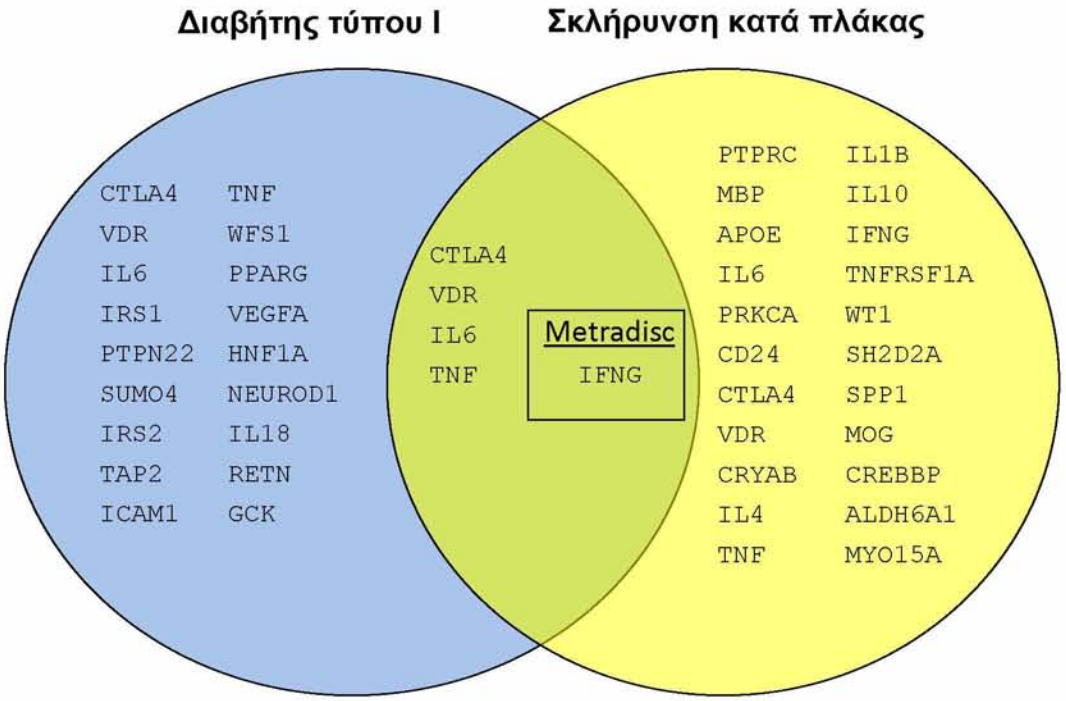
Εικόνα 4.3.1: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου I και Διαβήτη τύπου II.



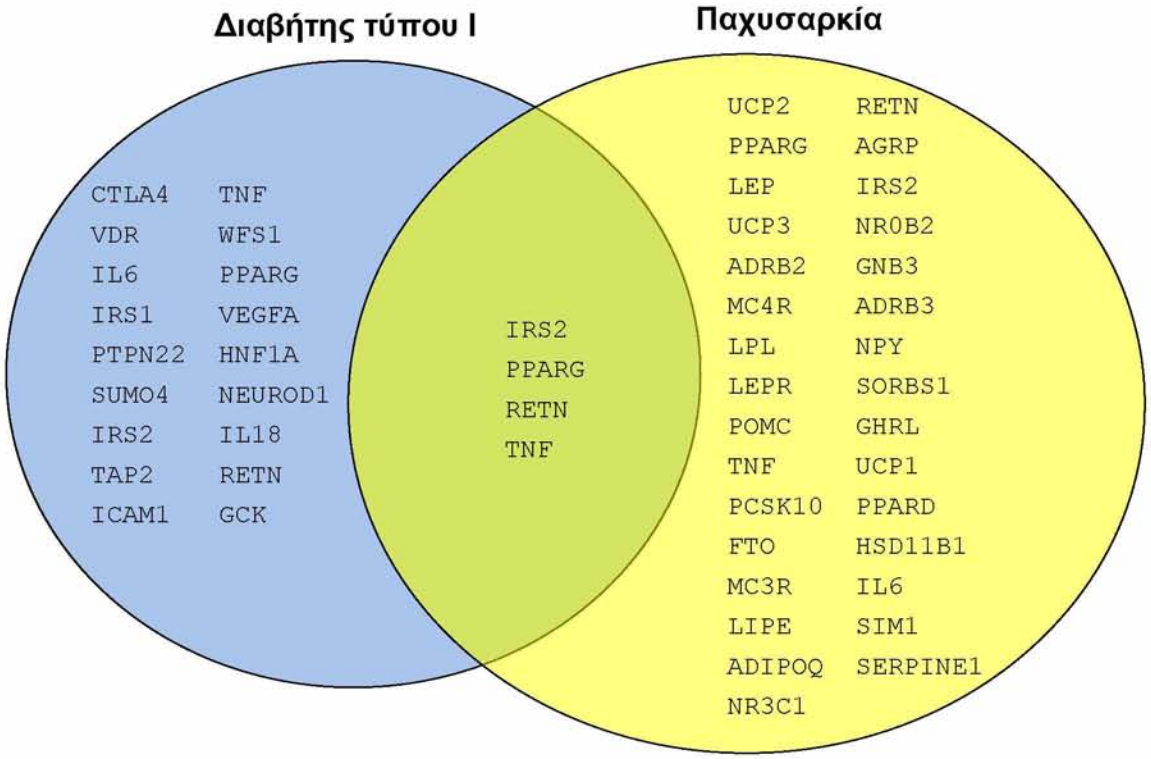
Εικόνα 4.3.2: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου Ι και Υπέρταση.



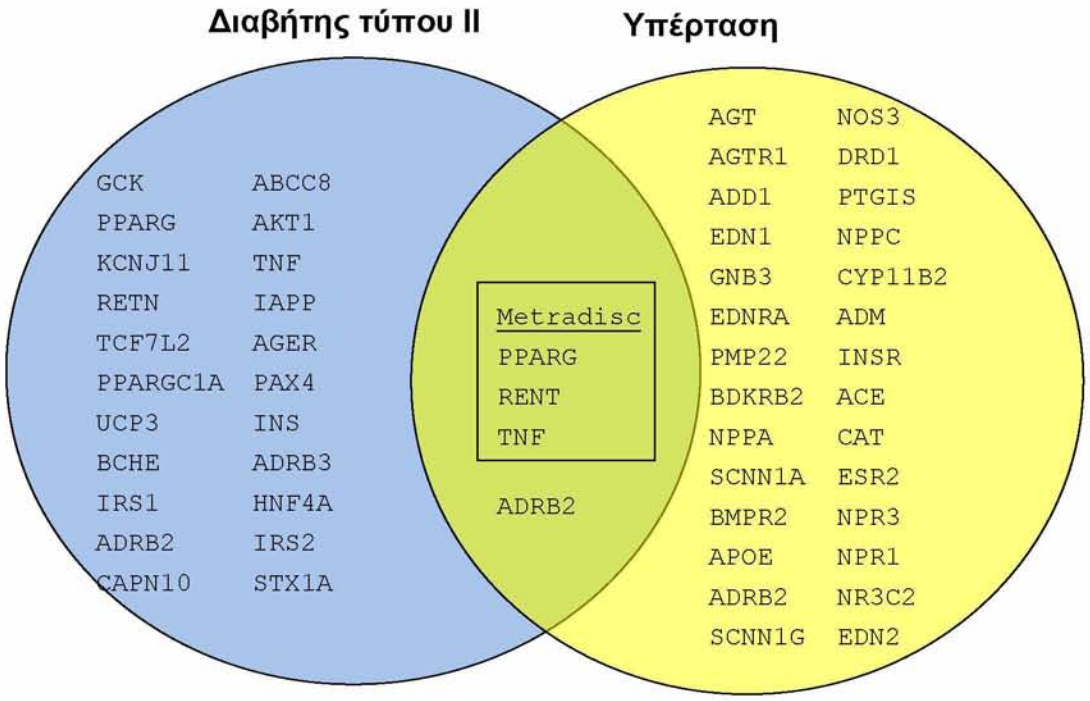
Εικόνα 4.3.3: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου Ι και Σκλήρυνση κατά πλάκα



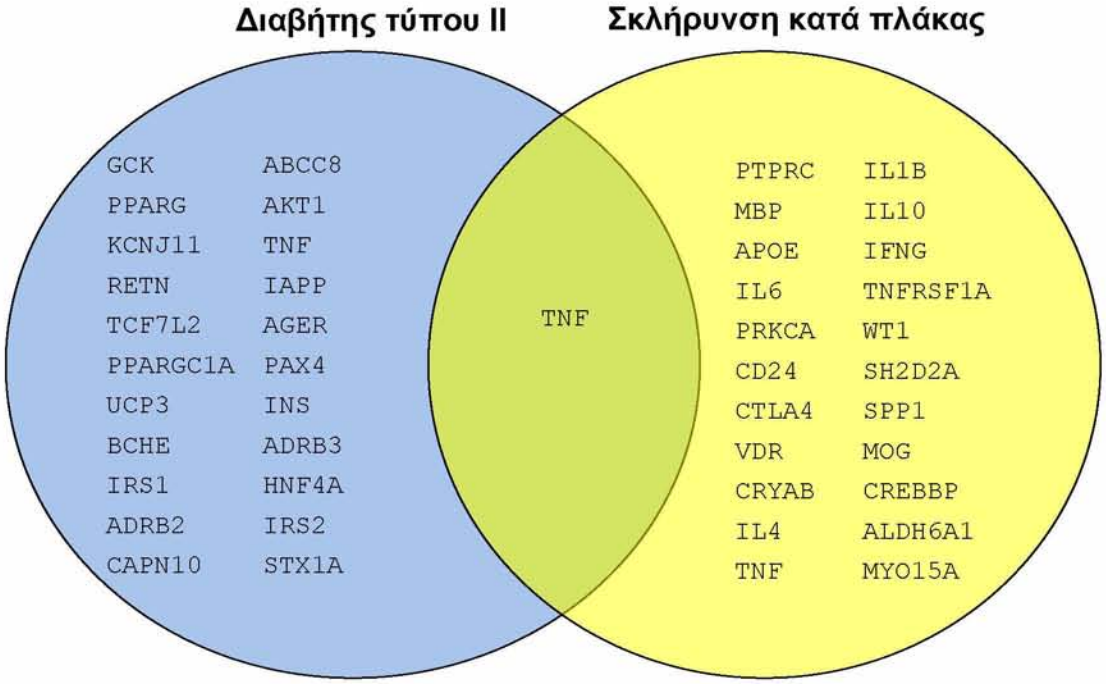
Εικόνα 4.3.4: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου Ι και Παχυσαρκία.



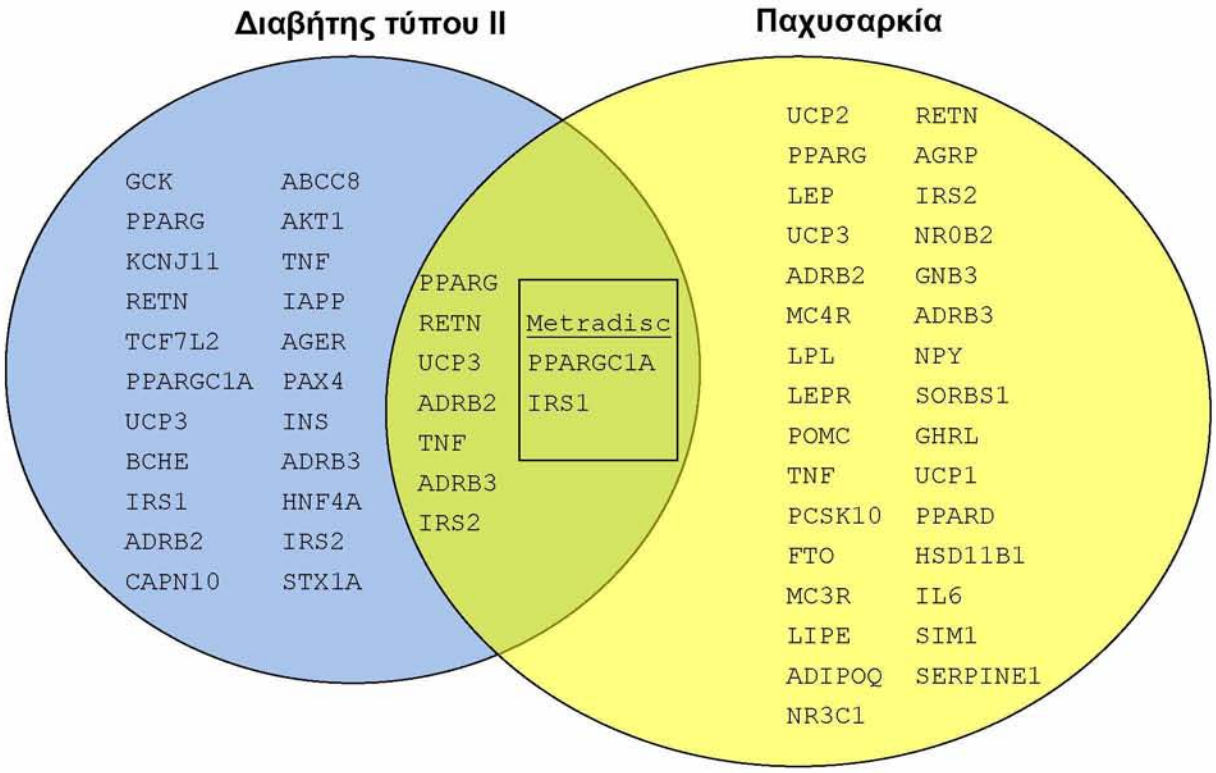
Εικόνα 4.3.5: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου ΙΙ και Υπέρταση.



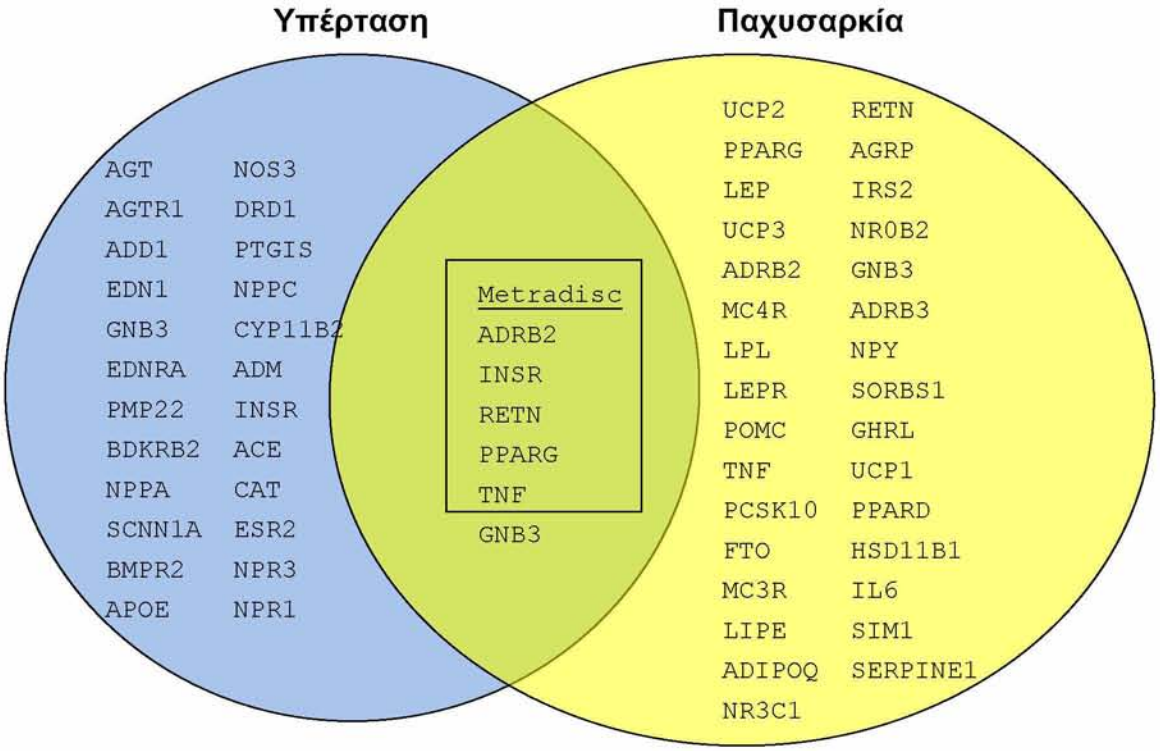
Εικόνα 4.3.6: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου ΙΙ και Σκλήρυνση κατά πλάκα.



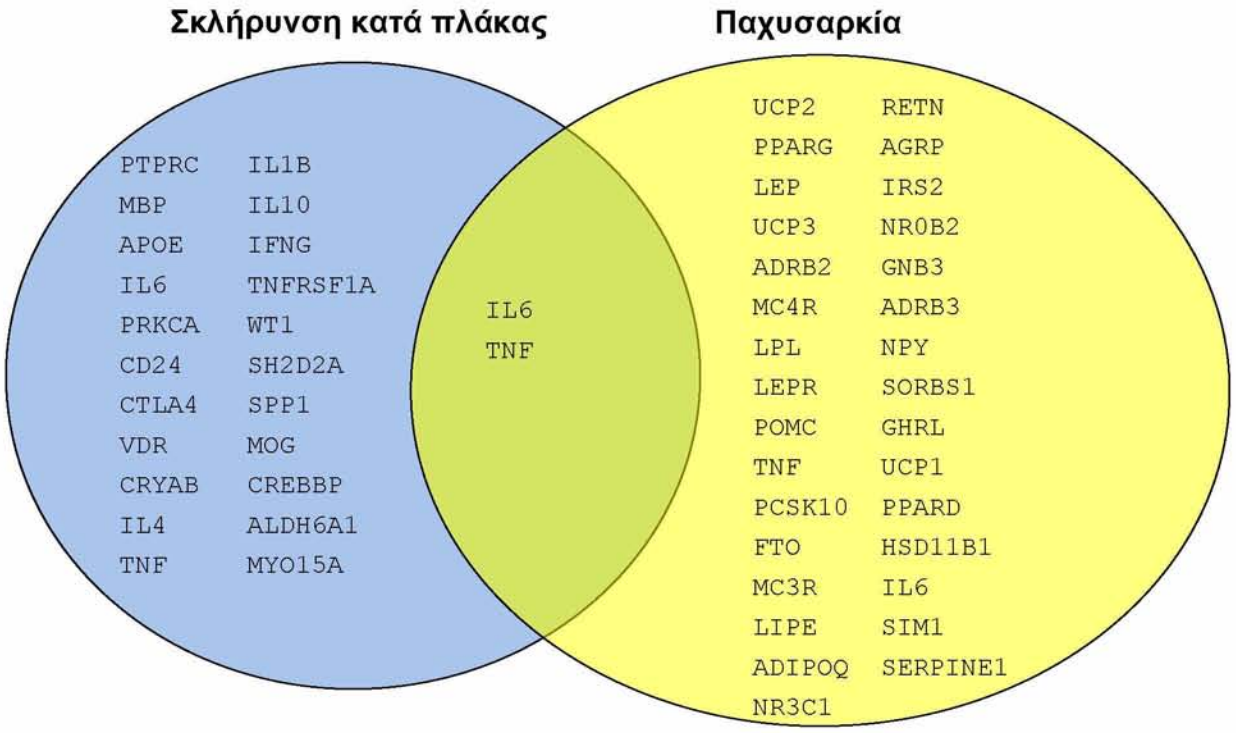
Εικόνα 4.3.7: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Διαβήτη τύπου ΙΙ και Παχυσαρκία.



Εικόνα 4.3.8: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Υπέρταση και Παχυσαρκία.



Εικόνα 4.3.9: Υψηλά υποψήφια γονίδια που βρέθηκε να σχετίζονται με τις ασθένειες Σκλήρυνση κατά πλάκα και Παχυσαρκία.



Από τα αποτελέσματα που παρουσιάζονται στις εικόνες 4.3.1 -4.3.9, αλλά και τα δεδομένα που παρουσιάζονται στον πίνακα 4.3.2 εύκολα μπορεί κανείς να ανιχνεύσει τις ασθένειες που πιθανός να συνδέονται μεταξύ τους. Είναι αξιοσημείωτο να σημειωθεί ότι με βάση τους διάφορους λόγους που έχουμε δημιουργήσει, βλέπουμε μία μεταβολή στα δεδομένα που βρίσκουμε όταν επεξεργαζόμαστε από τις πιο εδικές τιμές (σύνολο κοινών υψηλά υποψήφιων γονιδίων από τα αποτελέσματα RankProd) στις πιο γενικές τιμές (Σύνολο όλων των κοινών γονιδίων).

Η πιο πολύ συνδεδεμένη ασθένεια που παρουσιάζεται με βάση τα δεδομένα μας είναι ο Διαβήτης τύπου ΙΙ με την Παχυσαρκία. Αν και το σύνολο των γονιδίων που είχε συγκεντρωθεί και για τις δύο ασθένειες είναι το πιο χαμηλό (3436 γονίδια), το σύνολο όλων των κοινών γονιδίων που βρέθηκε να παρουσιάζονται στις δύο ασθένειες ήταν ένα από τα πιο μεγάλα σύνολα (1077 γονίδια). Επίσης ο συνδυασμός αυτών τον ασθενειών είχε τα περισσότερα κοινά υψηλά υποψήφια γονίδια. Από τα κοινά υψηλά υποψήφια γονίδια που χαρακτηρίζουν και τις δύο ασθένειες τα γονίδια UCP3[262], ADRB2-ADRB3[194, 263, 264] και TNF[265] βρέθηκε να σχετίζονται πάρα πολύ και με τις δύο ασθένειες. Τα γονίδια PPARG[266] ,IRS2 –IRS1[195, 267] και PPARGC1A[268, 269] βρέθηκε να σχετίζονται με τουλάχιστο μία από τις δύο ασθένειες και σίγουρα αποτελούν μεγάλο ενδιαφέρων για κατανόηση της σχέσεις μεταξύ των δύο ασθενειών.

Με βάση τα αποτελέσματα που παρουσιάζονται στο κεφάλαιο 3.3.5 (σελίδα 163) από την εύρεση των διαφορικά εκφρασμένων γονιδίων που έχουν οι δύο αυτές ασθένειες, ένα από τα κοινά υψηλά γονίδια που ανιχνεύουμε να σχετίζεται και με τις δύο ασθένειες το ADRB2 φαίνεται να σχετίζεται περισσότερο με τον Διαβήτη τύπου ΙΙ αν και βρέθηκε να σχετίζεται και με την Παχυσαρκία. Από τις δημοσιευμένες έρευνες που υπάρχουν καταχωρημένες στις διάφορες βάσεις δεδομένων, πολύ σωστά ανιχνεύουμε την σύνδεση του Διαβήτη τύπου ΙΙ με την παχυσαρκία. Συγκεκριμένα στη βάση δεδομένων PubMed βρέθηκαν 637 δημοσιεύσεις που σχετίζουν τον Διαβήτη τύπου ΙΙ με την Παχυσαρκία[270-274]. Όλα αυτά αποδεικνύουν ότι και οι δύο ασθένειες μαζί αποτελούν γενικότερα μία ασθένεια πολυπαραγοντικής αιτιολογίας.

Πίνακας 4.3.2: Συγκριτικός πίνακας συνδυασμένων ασθενειών, δημιουργία ειδικών παραμέτρων όπως τους Λόγους Α,Β,Γ,Δ,Ε για καλύτερη κατανόηση των σχέσεων μεταξύ των ασθενειών, αλλά και κατά πόσο συνδέονται μεταξύ τους.

Ασθένειες	CGR	CGRM	CTG	DER	TGR	TGRM	TG	Λόγος Α	Λόγος Β	Λόγος Γ	Λόγος Δ	Λόγος Ε
Διαβήτης τύπου I / Διαβήτη τύπου II	7	7	1491	11	40	67	3606	1.57	5.71	9.57	2.41	515.1
Διαβήτης τύπου I / Υπέρταση	3	3	1122	22	46	80	3778	7.33	15.33	26.7	3.36	1259
Διαβήτης τύπου I / Σκλήρυνση κατά πλάκα	4	5	1000	9	40	76	4160	2.25	10	15.2	4.16	832
Διαβήτης τύπου I / Παχυσαρκία	4	4	1137	21	49	70	3768	5.25	12.2	17.5	3.31	942
Διαβήτης τύπου II / Υπέρταση	1	4	1051	26	50	92	3446	26	50	23	3.27	861
Διαβήτης τύπου II / Σκλήρυνση κατά πλάκα	1	1	858	20	44	87	3828	20	44	87	4.46	3828
Διαβήτης τύπου II / Παχυσαρκία	7	9	1077	21	53	81	3436	3	7.57	9	3.19	381
Παχυσαρκία / Σκλήρυνση κατά πλάκα	2	2	873	17	53	90	3990	8.5	26.5	45	5.57	1995
Υπέρταση / Παχυσαρκία	1	6	1076	34	59	86	3608	34	59	14.3	3.35	601

$Λόγος\ A = CGR\ προς\ DER$
$Λόγος\ B = CGR\ προς\ TGR$
$Λόγος\ Γ = CGRM\ προς\ TGRM$
$Λόγος\ Δ = CTG\ προς\ TG$
$Λόγος\ E = TG\ προς\ CGRM$

CGR= Κοινά Υψηλά γονίδια (top genes) RankProd.
CGRM= Κοινά Υψηλά γονίδια (top genes) RankProd-Metradisc.
DER= Διαφορικά εκφρασμένα (RankProd).
TGRM= Σύνολο υψηλών γονιδίων (top genes)σε Metradisc-RankProd.
TGR= Σύνολο υψηλών γονιδίων(top genes) σε RankProd.
CTG= Σύνολο κοινών γονιδίων
TG= Σύνολο όλων γονιδίων.

Ένας δεύτερος συνδυασμός συσχετισμένων ασθενειών που ανιχνεύεται με βάση τα δεδομένα μας είναι η σύνδεση του Διαβήτη τύπου ΙΙ με την Υπέρταση. Επίσης και σε αυτό τον συνδυασμό ασθενειών παρατηρούμε το σύνολο των γονιδίων που είχε συγκεντρωθεί και για τις δύο ασθένειες να είναι χαμηλό (3446 γονίδια), ενώ το σύνολο όλων των κοινών γονιδίων που βρέθηκε να παρουσιάζονται και στις δύο ασθένειες να είναι από τα πιο μεγάλα σύνολα (1051). Τα κοινά υψηλά υποψήφια γονίδια που χαρακτηρίζουν και τις δύο ασθένειες δεν είναι πολλά (τέσσερα γονίδια) αλλά από αυτά τα τρία γονίδια, ADRB2[194, 275], TNF[276, 277] και PPARG[173, 278] φαίνεται να σχετίζονται αρκετά και με τις δύο ασθένειες. Επίσης και σε αυτό τον συνδυασμό ασθενειών, με βάση τα αποτελέσματα που παρουσιάζονται στο κεφάλαιο 3.3.7 (σελίδα 167) από την εύρεση των διαφορικά εκφρασμένων γονιδίων που έχουν οι δύο αυτές ασθένειες, ένα από τα κοινά υψηλά γονίδια που ανιχνεύουμε να σχετίζεται και με τις δύο, το ADRB2 φαίνεται να σχετίζεται περισσότερο με τον Διαβήτη τύπου ΙΙ αν και βρέθηκε να σχετίζεται και με την Υπέρταση. Για το τέταρτο γονίδιο το RETN όπου δεν βρέθηκε καμία δημοσιευμένη έρευνα, μετά από απλή αναζήτηση για πληροφορίες που υπάρχουν γύρο από αυτό το γονίδιο, μόλις 3 δημοσιευμένες έρευνες βρέθηκαν. Άρα πιθανός για το εν λόγο γονίδιο χρειαζόμαστε έρευνα στο μέλλον από την επιστημονική κοινότητα, για να είμαστε σίγουροι κατά πόσο σχετίζεται η όχι με τις εν λόγο ασθένειες.

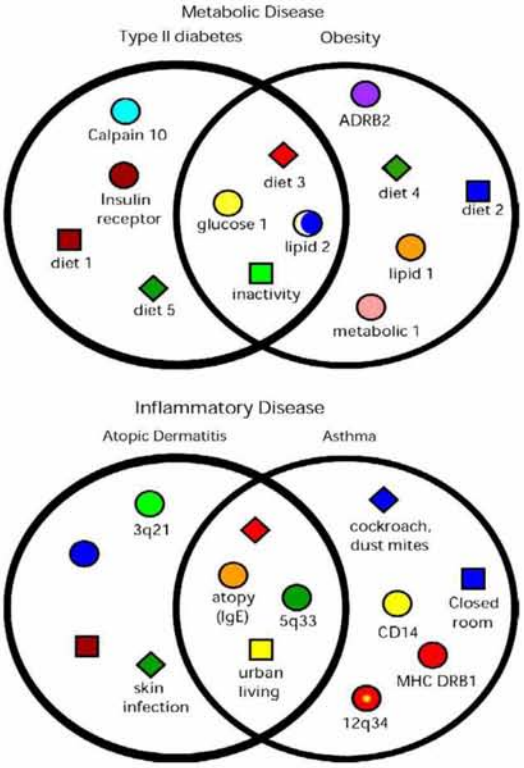
Όπως φαίνεται και στα αποτελέσματα που περιγράφουμε πιο πάνω και οι δύο ασθένειες Υπέρταση και Παχυσαρκία σχετίζονται με τον Διαβήτη τύπου ΙΙ. Άρα εύλογα κανείς θα υποψιαζόταν πιθανή συσχέτιση και της Υπέρτασης με την Παχυσαρκία. Με βάση τα αποτελέσματα του πίνακα 4.3.2 βλέπουμε ότι ο συσχετισμός μεταξύ των δύο αυτών ασθενειών είναι πολύ μεγάλος και μάλιστα παρουσιάζουν έξι κοινά υψηλά υποψήφια γονίδια. Από αυτά τα γονίδια, το γονίδιο GNB3[202] το οποίο ανιχνεύουμε να σχετίζεται υψηλά και με τις δύο ασθένειες (Υπέρταση – Παχυσαρκία) έχει συσχετιστεί να συνδέεται και με τις δύο αυτές ασθένειες με μεθόδους μετα-ανάλυσης από τον επίκουρο καθηγητή του Πανεπιστημίου Στερεάς Ελλάδος κ. Μπάγκο Παντελή. Αξιοσημείωτο είναι ότι το γονίδιο PPARG παρουσιάζεται υψηλά υποψήφιο και στις τρείς ασθένειες (Διαβήτη τύπου ΙΙ, Υπέρτασης και Παχυσαρκίας). Μετά από έρευνα για πιθανή σύνδεση και των τριών αυτών ασθενειών ως γενικότερα μία ασθένεια πολυπαραγοντικής αιτιολογίας, ανακαλύψαμε ότι και οι τρείς ασθένειες που βρήκαμε να σχετίζονται μεταξύ τους και να αποτελούν το μεταβολικό σύνδρομο[279].

Ένας τρίτος συνδυασμός ασθενειών που παρουσιάζουν μεγάλο βαθμό σύνδεσης είναι η ασθένεια του Διαβήτη τύπου ΙΙ με την ασθένεια του Διαβήτη τύπου Ι. Στις δύο αυτές ασθένειες θα πρέπει να είμαστε πολύ επιφυλακτική αφού και οι δύο ασθένειες περιέχουν πολλές κοινές λέξεις κλειδιά, και ίσος το όλο αποτέλεσμα να μην είναι και τόσο αξιόπιστο. Πέρα από αυτό, βλέπουμε ότι η δύο αυτές ασθένειες είχαν ένα από τα πιο μικρά σύνολά με γονίδια (3606) και πέτυχαν το πιο μεγάλο σύνολο από κοινά γονίδια (1491). Δεν μπορούμε να πούμε, ότι το πιο πάνω εύρημα οφείλεται σε λανθασμένο χαρακτηρισμό των γονιδίων, αφού με βάση τα επτά υψηλά υποψήφια γονίδια που ανιχνεύτηκαν να συνδέονται και με τις δύο ασθένειες βλέπουμε ότι τα πλείστα από αυτά IL6[158, 280], PPARG - IRS1[281, 282] έχει αποδεικτή να σχετίζονται άμεσα και στις δύο ασθένειες. Σε αυτή την περίπτωση θα πρέπει να αναμένουμε περισσότεροι έρευνα για να πούμε με σιγουριά ότι ο Διαβήτης τύπου Ι συνδέεται ή όχι με τον διαβήτη τύπου ΙΙ.

Τις τελευταίες δύο δεκαετίες παρατηρήθηκε μια έκρηξη στον εντοπισμό σε μεγάλο βαθμό των γονιδίων που συνδέονται με ασθένειες του Μέντελ. Περίπου 1.200 γονίδια έχουν χαρακτηριστεί και έχει κατανοηθεί η βάση της μοριακής γενετικής ασθένειας τους. Οι αρχές που απορρέουν από τις επιτυχίες αυτές θα πρέπει να εφαρμοστούν τώρα ως νέες στρατηγικές με στόχο την εύρεση των πιο ασαφών γονιδίων που αποτελούν τη βάση για την εξερεύνηση ασθενειών φαινοτυπικά πολυπαραγοντικής αιτιολογίας.

Με βάση έρευνες που έχουν διεξαχθεί, οι κοινές παραλλαγές ασθενειών μπορεί να αποτελούν υποθέσεις για μια ασθένεια πολυπαραγοντικής αιτιολογίας[260, 283, 284]. Τα επικαλυπτόμενα συνδεδεμένα γονίδια και οι σύνθετες γενετικές συσχετίσεις για το ίδιο γονίδιο ή για τα ίδια αλληλόμορφα, μαζί με τις κλινικά συναφείς διαταραχές δεν είναι σπάνιο εύρημα. Με βάση την προέκταση της υπόθεσης, κοινές παραλλαγές / ασθένειες πολυπαραγοντικής αιτιολογίας ορίζονται ως τα κοινά αλληλόμορφα που συμβάλλουν σε μία δεδομένη ασθένεια κάτω από ορισμένο συνδυασμό αλληλοσχετιζόμενων γονιδίων και περιβαλλοντικών συνθηκών και μπορεί να ενεργούν σε άλλα γενετικά υπόβαθρα επηρεαζόμενα από άλλους περιβαλλοντικούς παράγοντες και συνεπώς να έχουμε διαφορετικά ενδεχόμενα συνδεδεμένα κλινικά αποτελέσματα. Το Εικόνα 4.3.3 σκιαγραφεί ένα απλό γενικό μοντέλο όπου οι γενετικοί και περιβαλλοντικοί παράγοντες μοιράζονται μεταξύ δύο ασθενειών.

Εικόνα 4.3.3: Μοριακό μοντέλο των γενετικών και περιβαλλοντικών επικαλυπτόμενων παραγόντων σε κοινού τύπου ασθένειες.



Άρα η διαδικασία αυτή είναι πολύ καλύτερη από τις κλασικές μεθόδους σκοραρίσματος των γονιδίων με βάση το p-value. Η μέθοδος σκοραρίσματος των γονιδίων με βάση το p-value ελέγχει ταυτόχρονα 100-500.000 γονίδια. Αν πάρουμε ένα p-value 5% τότε σε βρίσκουμε στατιστικός σημαντικά π.χ. 2.000. Υπάρχει όμως πρόβλημα για αυξημένο σφάλμα τύπου 1 (πολλαπλές συγκρίσεις). Αν κάνουμε διόρθωση του p-value τότε από τα 2.000 απομένουν πολύ λίγα γονίδια (200) και χάνουμε την σημαντική πληροφορία. Με την μέθοδο που παρουσιάζουμε στην παρούσα εργασία πρώτα εκτελούμε ένα φιλτράρισμα των δεδομένων που παράγουμε από πολλές βιολογικές παραμέτρους και μετά προβαίνουμε σε στατιστικές αναλύσεις.

- Είναι εφικτό να συνδυαστούν διαφορετικού τύπου βιολογικά δεδομένα που παράγονται μέσα από διαφορετικές μεθόδους.
- (Consensus learning) .
- Πρώτη φορά εφαρμόζεται στην πράξη ο συνδυασμός διαφορετικού τύπου δεδομένων που παράγονται από τεχνικές 9 διαφορετικών αλγόριθμων. Έχει δοκιμαστεί και στο παρελθόν αλλά κάνοντας χρήση μόνο 3 τεχνικών.

Άρα αφού θα έχουμε αξιόπιστα δεδομένα και τα 2000 γονίδια που θα μας παρουσιάσει ως υψηλά γονίδια (top genes) θα είναι πιθανά υποψήφια και δεν θα χρειάζονται διόρθωση. Σκοπός είναι η υλοποίηση της διαδικασίας αυτής και να κατασκευαστή μία αυτόματη μηχανή που να εκτελεί όλες τις διαδικασίες που αναφέραμε.

Προβλήματα:

- Τρόπος εισαγωγής Λέξεων κλειδιών, αφού ορίζονται διαφορετικά σε κάθε αλγόριθμο αναζήτησης. Θα πρέπει να εκτελούνται αναζητήσεις με όλα τα συνώνυμα.

-Τρόπος διαβάσματος αποτελεσμάτων αλλά και συνδυασμός τους. (Υλοποιείται στην παρούσα εργασία)

Δεν υπάρχει ξεκάθαρη απάντηση στο πια μέθοδος, από τις υπάρχουσες μεθόδους που υπάρχουν μέχρι σήμερα είναι η καλύτερη και για αυτό τις συνδυάζουμε. Με βάση τα αποτελέσματα που είχαμε, φαίνεται να επιβεβαιώνεται σε πολύ μεγάλο βαθμό η προβλεπτική ικανότητα της όλης διαδικασίας αφού ήταν ικανή να προβλέψει γονίδια που είναι πλέον γνωστά ότι σχετίζονται με τις ασθένειες που αναζητούσαμε αλλά και γονίδια που δεν είναι ακόμη γνωστό να σχετίζονται με τις ασθένειες που αναζητούσαμε. Ανιχνεύουμε περισσότερα πιθανά υποψήφια γονίδια.

ΚΕΦΑΛΑΙΟ 5: ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Hong, F., et al., *RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.* Bioinformatics, 2006. **22**(22): p. 2825-7.
2. Zintzaras, E. and J.P. Ioannidis, *Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays.* Comput Biol Chem, 2008. **32**(1): p. 38-46.
3. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
4. Freudenberg, J. and P. Propping, *A similarity-based method for genome-wide prediction of disease-relevant human genes.* Bioinformatics, 2002. **18 Suppl 2**: p. S110-5.
5. Turner, F.S., D.R. Clutterbuck, and C.A. Semple, *POCUS: mining genomic sequence annotation to predict disease genes.* Genome Biol, 2003. **4**(11): p. R75.
6. Franke, L., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011-25.
7. George, R.A., et al., *Analysis of protein sequence and interaction data for candidate disease gene prediction.* Nucleic Acids Res, 2006. **34**(19): p. e130.
8. Rossi, S., et al., *TOM: a web-based integrated approach for identification of candidate disease genes.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W285-92.
9. Aerts, S., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
10. Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. **24**(1): p. 55-62.
11. Lage, K., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.
12. Bader, G.D., et al., *BIND--The Biomolecular Interaction Network Database.* Nucleic Acids Res, 2001. **29**(1): p. 242-5.
13. Radivojac, P., et al., *An integrated approach to inferring gene-disease associations in humans.* Proteins, 2008. **72**(3): p. 1030-7.
14. Risch, N.J., *Searching for genetic determinants in the new millennium.* Nature, 2000. **405**(6788): p. 847-56.
15. Glazier, A.M., J.H. Nadeau, and T.J. Aitman, *Finding genes that underlie complex traits.* Science, 2002. **298**(5602): p. 2345-9.
16. Mushegian, A.R., et al., *Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs.* Proc Natl Acad Sci U S A, 1997. **94**(11): p. 5831-6.
17. Jimenez-Sanchez, G., B. Childs, and D. Valle, *Human disease genes.* Nature, 2001. **409**(6822): p. 853-5.
18. Karlin, S., et al., *Amino acid runs in eukaryotic proteomes and disease associations.* Proc Natl Acad Sci U S A, 2002. **99**(1): p. 333-8.
19. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.* Nucleic Acids Res, 2003. **31**(13): p. 3692-7.
20. Zhang, C., et al., *Parallelization of multicategory support vector machines (PMC-SVM) for classifying microarray data.* BMC Bioinformatics, 2006. **7 Suppl 4**: p. S15.

21. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
22. Tiffin, N., et al., *Integration of text- and data-mining using ontologies successfully selects disease gene candidates.* Nucleic Acids Res, 2005. **33**(5): p. 1544-52.
23. Quan, H., et al., *Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.* Med Care, 2005. **43**(11): p. 1130-9.
24. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2005. **33**(Database issue): p. D514-7.
25. *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D142-8.
26. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
27. Brown, K.R. and I. Jurisica, *Online predicted human interaction database.* Bioinformatics, 2005. **21**(9): p. 2076-82.
28. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.
29. T., J., *A support vector method for multivariate performance measures.* Proceedings of the 22nd International Conference on Machine Learning (ICML). Bonn, Germany, 2005: p. 377–384.
30. Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
31. Goh, K.I., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
32. Yu, W., et al., *Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases.* BMC Bioinformatics, 2008. **9**: p. 528.
33. Guttmacher, A.E. and F.S. Collins, *Realizing the promise of genomics in biomedical research.* JAMA, 2005. **294**(11): p. 1399-402.
34. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
35. Khoury, M.J., et al., *The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?* Genet Med, 2007. **9**(10): p. 665-74.
36. Lin, B.K., et al., *Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database.* Am J Epidemiol, 2006. **164**(1): p. 1-4.
37. *NCBI.* National Center for Biotechnology Information, U.S. National Library of Medicine 2010: p. http://www.ncbi.nlm.nih.gov/guide/.
38. Hanson, E.H., G. Imperatore, and W. Burke, *HFE gene and hereditary hemochromatosis: a HuGE review. Human Genome Epidemiology.* Am J Epidemiol, 2001. **154**(3): p. 193-206.
39. *PubMed.* National Center for Biotechnology Information, U.S. National Library of Medicine
2010: p. http://www.ncbi.nlm.nih.gov/pubmed.

40. Adie, E.A., et al., *Speeding disease gene discovery by sequence based candidate prioritization.* BMC Bioinformatics, 2005. **6**: p. 55.
41. Pallen, M., B. Wren, and J. Parkhill, *'Going wrong with confidence': misleading sequence analyses of CiaB and clpX.* Mol Microbiol, 1999. **34**(1): p. 195.
42. Smith, N.G. and A. Eyre-Walker, *Human disease genes: patterns and predictions.* Gene, 2003. **318**: p. 169-75.
43. Kapetanovic, I.M., S. Rosenfeld, and G. Izmirlian, *Overview of commonly used bioinformatics methods and their applications.* Ann N Y Acad Sci, 2004. **1020**: p. 10-21.
44. Hammond, M.P. and E. Birney, *Genome information resources - developments at Ensembl.* Trends Genet, 2004. **20**(6): p. 268-72.
45. Winter, E.E., L. Goodstadt, and C.P. Ponting, *Elevated rates of protein secretion, evolution, and disease among tissue-specific genes.* Genome Res, 2004. **14**(1): p. 54-61.
46. Frank, E., et al., *Data mining in bioinformatics using Weka.* Bioinformatics, 2004. **20**(15): p. 2479-81.
47. Freund Y, M.L., *The Alternating Decision Tree Learning Algorithm.* Proceedings of the Sixteenth International Conference on Machine Learning: p. http://cseweb.ucsd.edu/~yfreund/papers/atrees.pdf.
48. Zhu, M., et al., *The K-nearest neighbor algorithm predicted rehabilitation potential better than current Clinical Assessment Protocol.* J Clin Epidemiol, 2007. **60**(10): p. 1015-21.
49. Brown, L.E., I. Tsamardinos, and C.F. Aliferis, *A novel algorithm for scalable and accurate Bayesian network learning.* Stud Health Technol Inform, 2004. **107**(Pt 1): p. 711-5.
50. Tanguay, R.L. and D.R. Gallie, *Translational efficiency is regulated by the length of the 3' untranslated region.* Mol Cell Biol, 1996. **16**(1): p. 146-56.
51. Cooper, D.N., P.D. Stenson, and N.A. Chuzhanova, *The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms.* Curr Protoc Bioinformatics, 2006. **Chapter 1**: p. Unit 1 13.
52. Mulder, N.J., et al., *The InterPro Database, 2003 brings increased coverage and new features.* Nucleic Acids Res, 2003. **31**(1): p. 315-8.
53. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
54. Huang, H., et al., *Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes.* Genome Biol, 2004. **5**(7): p. R47.
55. Group, T.F., *Identifying human disease genes.* Bios Scientific Publishers , Human Molecular Genetics 1999: p. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hmg&part=A1867.
56. Becker, K.G., et al., *The genetic association database.* Nat Genet, 2004. **36**(5): p. 431-2.
57. Yue, P., E. Melamud, and J. Moult, *SNPs3D: candidate gene and SNP selection for association studies.* BMC Bioinformatics, 2006. **7**: p. 166.
58. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
59. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.

60. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.
61. *KEGG pathway database.* . 2010: p. http://www.genome.jp/kegg/.
62. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32**(Database issue): p. D497-501.
63. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
64. Lee, I., et al., *A probabilistic functional network of yeast genes.* Science, 2004. **306**(5701): p. 1555-8.
65. Tong, A.H., et al., *Global mapping of the yeast genetic interaction network.* Science, 2004. **303**(5659): p. 808-13.
66. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
67. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
68. Phizicky, E., et al., *Protein analysis on a proteomic scale.* Nature, 2003. **422**(6928): p. 208-15.
69. Lin, N., et al., *Information assessment on predicting protein-protein interactions.* BMC Bioinformatics, 2004. **5**: p. 154.
70. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.
71. Yue, P., Z. Li, and J. Moult, *Loss of protein structure stability as a major causative factor in monogenic disease.* J Mol Biol, 2005. **353**(2): p. 459-73.
72. Yue, P. and J. Moult, *Identification and analysis of deleterious human SNPs.* J Mol Biol, 2006. **356**(5): p. 1263-74.
73. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
74. Sunyaev, S., et al., *Prediction of deleterious human alleles.* Hum Mol Genet, 2001. **10**(6): p. 591-7.
75. Ramensky, V., P. Bork, and S. Sunyaev, *Human non-synonymous SNPs: server and survey.* Nucleic Acids Res, 2002. **30**(17): p. 3894-900.
76. Krishnan, V.G. and D.R. Westhead, *A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function.* Bioinformatics, 2003. **19**(17): p. 2199-209.
77. Reumers, J., et al., *SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D527-32.
78. Cai, Z., et al., *Bayesian approach to discovering pathogenic SNPs in conserved protein domains.* Hum Mutat, 2004. **24**(2): p. 178-84.
79. Saunders, C.T. and D. Baker, *Evaluation of structural and evolutionary contributions to deleterious mutation prediction.* J Mol Biol, 2002. **322**(4): p. 891-901.
80. Karchin, R., L. Kelly, and A. Sali, *Improving functional annotation of non-synonomous SNPs with information theory.* Pac Symp Biocomput, 2005: p. 397-408.
81. Mooney, S.D. and R.B. Altman, *MutDB: annotating human variation with functionally relevant data.* Bioinformatics, 2003. **19**(14): p. 1858-60.

82. Stitziel, N.O., et al., *topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association.* Nucleic Acids Res, 2004. **32**(Database issue): p. D520-2.
83. Cavallo, A. and A.C. Martin, *Mapping SNPs to protein sequence and structure data.* Bioinformatics, 2005. **21**(8): p. 1443-50.
84. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
85. Chen, R., et al., *FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease.* Genome Biol, 2008. **9**(12): p. R170.
86. Barrett, T., et al., *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.
87. Chen, R., L. Li, and A.J. Butte, *AILUN: reannotating gene expression data automatically.* Nat Methods, 2007. **4**(11): p. 879.
88. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
89. Blekhman, R., et al., *Natural selection on genes that underlie human disease susceptibility.* Curr Biol, 2008. **18**(12): p. 883-9.
90. Yoshida, Y., et al., *PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W147-52.
91. *MEDLINE.* National Institutes of Health , Health & Human Services, 2006: p. http://www.nlm.nih.gov/databases/databases_medline.html.
92. Kobayashi, N. and T. Toyoda, *Statistical search on the Semantic Web.* Bioinformatics, 2008. **24**(7): p. 1002-10.
93. Prud'hommeaux, E.a.S., A., *SPARQL Query Language for RDF.* W3C Recommendation 2008: p. http://www.w3.org/TR/rdf-sparql-query/.
94. Coletti, M.H. and H.L. Bleich, *Medical subject headings used to search the biomedical literature.* J Am Med Inform Assoc, 2001. **8**(4): p. 317-23.
95. *RIKEN* The Institute of Physical and Chemical Research, Japan: p. http://omicspace.riken.jp/acknwldgmnt.htm.
96. Makino, T. and T. Gojobori, *Evolution of protein-protein interaction network.* Genome Dyn, 2007. **3**: p. 13-29.
97. Cui, J., et al., *AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology.* Nucleic Acids Res, 2008. **36**(Database issue): p. D999-1008.
98. Obayashi, T., et al., *ATTED-II provides coexpressed gene networks for Arabidopsis.* Nucleic Acids Res, 2009. **37**(Database issue): p. D987-91.
99. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.
100. Blake, J.A., et al., *The Mouse Genome Database genotypes::phenotypes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D712-9.
101. Dwinell, M.R., et al., *The Rat Genome Database 2009: variation, ontologies and pathways.* Nucleic Acids Res, 2009. **37**(Database issue): p. D744-9.
102. Wain, H.M., et al., *Genew: the human gene nomenclature database.* Nucleic Acids Res, 2002. **30**(1): p. 169-71.

103. Swarbreck, D., et al., *The Arabidopsis Information Resource (TAIR): gene structure and function annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D1009-14.
104. Takahashi, H., et al., *Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry.* Anal Bioanal Chem, 2008. **391**(8): p. 2769-82.
105. van Driel, M.A., et al., *GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W758-61.
106. Adie, E.A., et al., *SUSPECTS: enabling fast and effective prioritization of positional candidates.* Bioinformatics, 2006. **22**(6): p. 773-4.
107. Seelow, D., J.M. Schwarz, and M. Schuelke, *GeneDistiller--distilling candidate genes from linkage intervals.* PLoS One, 2008. **3**(12): p. e3874.
108. Kohler, S., et al., *Walking the interactome for prioritization of candidate disease genes.* Am J Hum Genet, 2008. **82**(4): p. 949-58.
109. *GeneSniffer.* 2009: p. http://www.genesniffer.org.
110. Crans, G.G. and J.J. Shuster, *How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial.* Stat Med, 2008. **27**(18): p. 3598-611.
111. Hutz, J.E., et al., *CANDID: a flexible method for prioritizing candidate genes for complex human traits.* Genet Epidemiol, 2008. **32**(8): p. 779-90.
112. Karimpour-Fard, A., et al., *Cross-species cluster co-conservation: a new method for generating protein interaction networks.* Genome Biol, 2007. **8**(9): p. R185.
113. Lockhart, D.J. and E.A. Winzeler, *Genomics, gene expression and DNA arrays.* Nature, 2000. **405**(6788): p. 827-36.
114. Marchler-Bauer, A., et al., *CDD: a Conserved Domain Database for protein classification.* Nucleic Acids Res, 2005. **33**(Database issue): p. D192-6.
115. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2002. **30**(1): p. 276-80.
116. Letunic, I., et al., *SMART 4.0: towards genomic data integration.* Nucleic Acids Res, 2004. **32**(Database issue): p. D142-4.
117. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
118. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
119. Safran, M., et al., *GeneCards Version 3: the human gene integrator.* Database (Oxford), 2010. **2010**: p. baq020.
120. Rebhan, M., et al., *GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.* Bioinformatics, 1998. **14**(8): p. 656-64.
121. Fangxin Hong, B.W., *Bioconductor RankProd Package Vignette.* 2005: p. http://arabidopsis.org/info/expression/ATGenExpress.jsp.
122. Dessau, R.B. and C.B. Pipper, *["R"--project for statistical computing].* Ugeskr Laeger, 2008. **170**(5): p. 328-30.
123. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
124. Zintzaras, E. and J.P. Ioannidis, *HEGESMA: genome search meta-analysis and heterogeneity testing.* Bioinformatics, 2005. **21**(18): p. 3672-3.

125. Dhanasekaran, S.M., et al., *Delineation of prognostic biomarkers in prostate cancer.* Nature, 2001. **412**(6849): p. 822-6.
126. Zintzaras, E. and J.P. Ioannidis, *Heterogeneity testing in meta-analysis of genome searches.* Genet Epidemiol, 2005. **28**(2): p. 123-37.
127. Thornblad, T.A., et al., *Prioritization of positional candidate genes using multiple web-based software tools.* Twin Res Hum Genet, 2007. **10**(6): p. 861-70.
128. Hong, F. and R. Breitling, *A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.* Bioinformatics, 2008. **24**(3): p. 374-82.
129. Esteva, F.J., et al., *PTEN, PIK3CA, p-AKT, and p-p70S6K Status. Association with Trastuzumab Response and Survival in Patients with HER2-Positive Metastatic Breast Cancer.* Am J Pathol, 2010.
130. Fabi, A., et al., *Clinical significance of PTEN and p-Akt co-expression in HER2-positive metastatic breast cancer patients treated with trastuzumab-based therapies.* Oncology, 2010. **78**(2): p. 141-9.
131. Long, J.R., et al., *Genetic polymorphisms of the CYP19A1 gene and breast cancer survival.* Cancer Epidemiol Biomarkers Prev, 2006. **15**(11): p. 2115-22.
132. Cai, Q., et al., *Haplotype analyses of CYP19A1 gene variants and breast cancer risk: results from the Shanghai Breast Cancer Study.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(1): p. 27-32.
133. Gao, L.B., et al., *RAD51 135G/C polymorphism and breast cancer risk: a meta-analysis from 21 studies.* Breast Cancer Res Treat, 2010.
134. Zhou, G.W., et al., *RAD51 135G>C polymorphism and breast cancer risk: a meta-analysis.* Breast Cancer Res Treat, 2010.
135. Weischer, M., et al., *CHEK2*1100delC genotyping for clinical assessment of breast cancer risk: meta-analyses of 26,000 patient cases and 27,000 controls.* J Clin Oncol, 2008. **26**(4): p. 542-8.
136. Guenard, F., et al., *Evaluation of the contribution of the three breast cancer susceptibility genes CHEK2, STK11, and PALB2 in non-BRCA1/2 French Canadian families with high risk of breast cancer.* Genet Test Mol Biomarkers, 2010. **14**(4): p. 515-26.
137. Lu, C., et al., *CCND1 G870A polymorphism contributes to breast cancer susceptibility: a meta-analysis.* Breast Cancer Res Treat, 2009. **116**(3): p. 571-5.
138. Onay, U.V., et al., *Combined effect of CCND1 and COMT polymorphisms and increased breast cancer risk.* BMC Cancer, 2008. **8**: p. 6.
139. Economopoulos, K.P. and T.N. Sergentanis, *Differential effects of MDM2 SNP309 polymorphism on breast cancer risk along with race: a meta-analysis.* Breast Cancer Res Treat, 2010. **120**(1): p. 211-6.
140. Elsberger, B., et al., *Breast cancer patients' clinical outcome measures are associated with Src kinase family member expression.* Br J Cancer, 2010. **103**(6): p. 899-909.
141. Chen, Y., et al., *Combined Src and ER blockade impairs human breast cancer proliferation in vitro and in vivo.* Breast Cancer Res Treat, 2010.
142. Jakubowska, A., et al., *CDH1 gene mutations do not contribute in hereditary diffuse gastric cancer in Poland.* Fam Cancer, 2010.
143. Wong, C.M., et al., *Quantitative analysis of promoter methylation in exfoliated epithelial cells isolated from breast milk of healthy women.* Epigenetics, 2010. **5**(7).

144. Deming, S.L., et al., *C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance.* Br J Cancer, 2000. **83**(12): p. 1688-95.
145. Alles, M.C., et al., *Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup.* PLoS One, 2009. **4**(3): p. e4710.
146. Mangoni, M., et al., *Association between Genetic Polymorphisms in the XRCC1, XRCC3, XPD, GSTM1, GSTT1, MSH2, MLH1, MSH3, and MGMT Genes and Radiosensitivity in Breast Cancer Patients.* Int J Radiat Oncol Biol Phys, 2010.
147. Qiu, L.X., et al., *XRCC3 5'-UTR and IVS5-14 polymorphisms and breast cancer susceptibility: a meta-analysis.* Breast Cancer Res Treat, 2010. **122**(2): p. 489-93.
148. Subik, K., et al., *The Expression Patterns of ER, PR, HER2, CK5/6, EGFR, Ki-67 and AR by Immunohistochemical Analysis in Breast Cancer Cell Lines.* Breast Cancer (Auckl), 2010. **4**: p. 35-41.
149. Graham, T.R., et al., *Reciprocal regulation of ZEB1 and AR in triple negative breast cancer cells.* Breast Cancer Res Treat, 2010. **123**(1): p. 139-47.
150. Sabatier, R., et al., *BARD1 homozygous deletion, a possible alternative to BRCA1 mutation in basal breast cancer.* Genes Chromosomes Cancer, 2010.
151. De Brakeleer, S., et al., *Cancer predisposing missense and protein truncating BARD1 mutations in non-BRCA1 or BRCA2 breast cancer families.* Hum Mutat, 2010. **31**(3): p. E1175-85.
152. Armanious, H., et al., *STAT3 upregulates the protein expression and transcriptional activity of beta-catenin in breast cancer.* Int J Clin Exp Pathol, 2010. **3**(7): p. 654-64.
153. Lin, L., et al., *Novel STAT3 phosphorylation inhibitors exhibit potent growth-suppressive activity in pancreatic and breast cancer cells.* Cancer Res, 2010. **70**(6): p. 2445-54.
154. Ma, Y., et al., *No significant association between the TP53 codon 72 polymorphism and breast cancer risk: a meta-analysis of 21 studies involving 24,063 subjects.* Breast Cancer Res Treat, 2010.
155. Hu, Z., et al., *Three common TP53 polymorphisms in susceptibility to breast cancer, evidence from meta-analysis.* Breast Cancer Res Treat, 2010. **120**(3): p. 705-14.
156. Burwinkel, B., et al., *Association of NCOA3 polymorphisms with breast cancer risk.* Clin Cancer Res, 2005. **11**(6): p. 2169-74.
157. Wilkening, S., et al., *Polyglutamine repeat length in the NCOA3 does not affect risk in familial breast cancer.* Cancer Epidemiol Biomarkers Prev, 2005. **14**(1): p. 291-2.
158. Qi, L., et al., *Genetic variation in IL6 gene and type 2 diabetes: tagging-SNP haplotype analysis in large-scale case-control study and meta-analysis.* Hum Mol Genet, 2006. **15**(11): p. 1914-20.
159. Bouhaha, R., et al., *Study of TNFalpha -308G/A and IL6 -174G/C polymorphisms in type 2 diabetes and obesity risk in the Tunisian population.* Clin Biochem, 2010. **43**(6): p. 549-52.
160. Brann, B.S.t., et al., *Asymmetric growth of the lateral cerebral ventricle in infants with posthemorrhagic ventricular dilation.* J Pediatr, 1991. **118**(1): p. 108-12.
161. Cyganek, K., et al., *Clinical risk factors and the role of VDR gene polymorphisms in diabetic retinopathy in Polish type 2 diabetes patients.* Acta Diabetol, 2006. **43**(4): p. 114-9.

162. Feng, R.N., Y. Li, and C.H. Sun, *TNF 308 G/A polymorphism and type 1 diabetes: a meta-analysis.* Diabetes Res Clin Pract, 2009. **85**(1): p. e4-7.
163. Javor, J., et al., *Polymorphisms in the genes encoding TGF-beta1, TNF-alpha, and IL-6 show association with type 1 diabetes mellitus in the Slovak population.* Arch Immunol Ther Exp (Warsz), 2010. **58**(5): p. 385-93.
164. Al-Kateb, H., et al., *Multiple variants in vascular endothelial growth factor (VEGFA) are risk factors for time to severe retinopathy in type 1 diabetes: the DCCT/EDIC genetics study.* Diabetes, 2007. **56**(8): p. 2161-8.
165. Rau, H., et al., *The codon 17 polymorphism of the CTLA4 gene in type 2 diabetes mellitus.* J Clin Endocrinol Metab, 2001. **86**(2): p. 653-5.
166. Cooper, J.D., et al., *The candidate genes TAF5L, TCF7, PDCD1, IL6 and ICAM1 cannot be excluded from having effects in type 1 diabetes.* BMC Med Genet, 2007. **8**: p. 71.
167. Larsen, Z.M., et al., *Evidence for linkage on chromosome 4p16.1 in Type 1 diabetes Danish families and complete mutation scanning of the WFS1 (Wolframin) gene.* Diabet Med, 2004. **21**(3): p. 218-22.
168. Porksen, S., et al., *Variation within the PPARG gene is associated with residual beta-cell function and glycemic control in children and adolescents during the first year of clinical type 1 diabetes.* Pediatr Diabetes, 2008. **9**(4 Pt 1): p. 297-302.
169. Boraska, V., et al., *NeuroD1 gene and interleukin-18 gene polymorphisms in type 1 diabetes in Dalmatian population of Southern Croatia.* Croat Med J, 2006. **47**(4): p. 571-8.
170. Oliveira, C.S., et al., *The Ala45Thr polymorphism of NEUROD1 is associated with type 1 diabetes in Brazilian women.* Diabetes Metab, 2005. **31**(6): p. 599-602.
171. Khalil, R., et al., *Screening of mutations in the GCK gene in Jordanian maturity-onset diabetes of the young type 2 (MODY2) patients.* Genet Mol Res, 2009. **8**(2): p. 500-6.
172. Reiling, E., et al., *Combined effects of single-nucleotide polymorphisms in GCK, GCKR, G6PC2 and MTNR1B on fasting plasma glucose and type 2 diabetes risk.* Diabetologia, 2009. **52**(9): p. 1866-70.
173. Gupta, V. and S. Ebrahim, *Comment on: Chauhan et al. (2010) Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians. Diabetes;59:2068-2074.* Diabetes, 2010. **59**(9): p. e15; author reply e16.
174. Chauhan, G., et al., *Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians.* Diabetes, 2010. **59**(8): p. 2068-74.
175. Zhou, D., et al., *The E23K variation in the KCNJ11 gene is associated with type 2 diabetes in Chinese and East Asian population.* J Hum Genet, 2009. **54**(7): p. 433-5.
176. Sakamoto, Y., et al., *SNPs in the KCNJ11-ABCC8 gene locus are associated with type 2 diabetes and blood pressure levels in the Japanese population.* J Hum Genet, 2007. **52**(10): p. 781-93.
177. Barroso, I., et al., *Meta-analysis of the Gly482Ser variant in PPARGC1A in type 2 diabetes and related phenotypes.* Diabetologia, 2006. **49**(3): p. 501-5.

178. Morini, E., et al., *IRS1 G972R polymorphism and type 2 diabetes: a paradigm for the difficult ascertainment of the contribution to disease susceptibility of 'low-frequency-low-risk' variants.* Diabetologia, 2009. **52**(9): p. 1852-7.
179. Zeggini, E., et al., *Association studies of insulin receptor substrate 1 gene (IRS1) variants in type 2 diabetes samples enriched for family history and early age of onset.* Diabetes, 2004. **53**(12): p. 3319-22.
180. Ridderstrale, M. and E. Nilsson, *Type 2 diabetes candidate gene CAPN10: first, but not last.* Curr Hypertens Rep, 2008. **10**(1): p. 19-24.
181. Vander Molen, J., et al., *Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants.* Am J Hum Genet, 2005. **76**(4): p. 548-60.
182. Gloyn, A.L., et al., *Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes.* Diabetes, 2003. **52**(2): p. 568-72.
183. Boraska, V., et al., *Large-scale association analysis of TNF/LTA gene region polymorphisms in type 2 diabetes.* BMC Med Genet, 2010. **11**: p. 69.
184. Feng, R., et al., *Lack of association between TNF 238 G/A polymorphism and type 2 diabetes: a meta-analysis.* Acta Diabetol, 2009. **46**(4): p. 339-43.
185. Johansson, S., et al., *Studies in 3,523 Norwegians and meta-analysis in 11,571 subjects indicate that variants in the hepatocyte nuclear factor 4 alpha (HNF4A) P2 region are associated with type 2 diabetes in Scandinavians.* Diabetes, 2007. **56**(12): p. 3112-7.
186. Luo, Y., et al., *Meta-analysis of the association between SNPs in TCF7L2 and type 2 diabetes in East Asian population.* Diabetes Res Clin Pract, 2009. **85**(2): p. 139-46.
187. Tong, Y., et al., *Association between TCF7L2 gene polymorphisms and susceptibility to type 2 diabetes mellitus: a large Human Genome Epidemiology (HuGE) review and meta-analysis.* BMC Med Genet, 2009. **10**: p. 15.
188. Rudofsky, G., Jr., et al., *Promoter polymorphisms of UCP1, UCP2, and UCP3 are not associated with diabetic microvascular complications in type 2 diabetes.* Horm Metab Res, 2007. **39**(4): p. 306-9.
189. Pinelli, M., et al., *Beta2-adrenergic receptor and UCP3 variants modulate the relationship between age and type 2 diabetes mellitus.* BMC Med Genet, 2006. **7**: p. 85.
190. Goulart, A.C., et al., *Polymorphisms in advanced glycosylation end product-specific receptor (AGER) gene, insulin resistance, and type 2 diabetes mellitus.* Clin Chim Acta, 2008. **398**(1-2): p. 95-8.
191. Tokuyama, Y., et al., *The Arg121Trp variant in PAX4 gene is associated with beta-cell dysfunction in Japanese subjects with type 2 diabetes mellitus.* Metabolism, 2006. **55**(2): p. 213-6.
192. Shimajiri, Y., et al., *A missense mutation of Pax4 gene (R121W) is associated with type 2 diabetes in Japanese.* Diabetes, 2001. **50**(12): p. 2864-9.
193. Yang, M., et al., *Effects of UCP2 -866 G/A and ADRB3 Trp64Arg on rosiglitazone response in Chinese patients with Type 2 diabetes.* Br J Clin Pharmacol, 2009. **68**(1): p. 14-22.
194. Kilpelainen, T.O., et al., *Interaction of single nucleotide polymorphisms in ADRB2, ADRB3, TNF, IL6, IGF1R, LIPC, LEPR, and GHRL with physical*

activity on the risk of type 2 diabetes mellitus and changes in characteristics of the metabolic syndrome: The Finnish Diabetes Prevention Study. Metabolism, 2008. **57**(3): p. 428-36.
195. Brady, M.J., *IRS2 takes center stage in the development of type 2 diabetes.* J Clin Invest, 2004. **114**(7): p. 886-8.
196. Mondry, A., et al., *Polymorphisms of the insertion / deletion ACE and M235T AGT genes and hypertension: surprising new findings and meta-analysis of data.* BMC Nephrol, 2005. **6**(1): p. 1.
197. Masi, M., F. Vivarelli, and V. Vecchi, *[Serum C3 levels at various pediatric ages in normal and pathologic conditions].* Minerva Pediatr, 1973. **25**(16): p. 731-5.
198. Girerd, B., et al., *Absence of influence of gender and BMPR2 mutation type on clinical phenotypes of pulmonary arterial hypertension.* Respir Res, 2010. **11**: p. 73.
199. Johri, S., G.H. Dunnington, and C.L. Vnencak-Jones, *A novel BMPR2 mutation associated with pulmonary arterial hypertension in an octogenarian.* Lung, 2010. **188**(4): p. 349-52.
200. Cheng, X. and G. Xu, *Association between aldosterone synthase CYP11B2 polymorphism and essential hypertension in Chinese: a meta-analysis.* Kidney Blood Press Res, 2009. **32**(2): p. 128-40.
201. Ramu, P., et al., *Gly460Trp polymorphism of the ADD1 gene and essential hypertension in an Indian population: A meta-analysis on hypertension risk.* Indian J Hum Genet, 2010. **16**(1): p. 8-15.
202. Bagos, P.G., et al., *The GNB3 C825T polymorphism and essential hypertension: a meta-analysis of 34 studies including 14,094 cases and 17,760 controls.* J Hypertens, 2007. **25**(3): p. 487-500.
203. Nie, S.J., et al., *Haplotype-based case-control study of the human AGTR1 gene and essential hypertension in Han Chinese subjects.* Clin Biochem, 2010. **43**(3): p. 253-8.
204. Chung, W.K., et al., *Polymorphism in the angiotensin II type 1 receptor (AGTR1) is associated with age at diagnosis in pulmonary arterial hypertension.* J Heart Lung Transplant, 2009. **28**(4): p. 373-9.
205. Lynch, A.I., et al., *Pharmacogenetic association of the NPPA T2238C genetic variant with cardiovascular disease outcomes in patients with hypertension.* JAMA, 2008. **299**(3): p. 296-307.
206. Benjafield, A.V., K. Katyk, and B.J. Morris, *Association of EDNRA, but not WNK4 or FKBP1B, polymorphisms with essential hypertension.* Clin Genet, 2003. **64**(5): p. 433-8.
207. Buday, A., et al., *Elevated systemic TGF-beta impairs aortic vasomotor function through activation of NADPH oxidase-driven superoxide production and leads to hypertension, myocardial remodeling, and increased plaque formation in apoE(-/-) mice.* Am J Physiol Heart Circ Physiol, 2010. **299**(2): p. H386-95.
208. Kester, M.I., et al., *Joint effect of hypertension and APOE genotype on CSF biomarkers for Alzheimer's disease.* J Alzheimers Dis, 2010. **20**(4): p. 1083-90.
209. Kingah, P.L., et al., *Association of NOS3 Glu298Asp SNP with hypertension and possible effect modification of dietary fat intake in the ARIC study.* Hypertens Res, 2010. **33**(2): p. 165-9.

210. Cruz-Gonzalez, I., et al., *Association between -T786C NOS3 polymorphism and resistant hypertension: a prospective cohort study.* BMC Cardiovasc Disord, 2009. **9**: p. 35.
211. Beige, J., et al., *Ethnic origin determines the impact of genetic variants in dopamine receptor gene (DRD1) concerning essential hypertension.* Am J Hypertens, 2004. **17**(12 Pt 1): p. 1184-7.
212. Yoshino, N., et al., *[Metastatic endometrial stromal sarcoma successfully treated by intra-arterial hypertension chemotherapy with CDDP and ADM].* Gan To Kagaku Ryoho, 1990. **17**(8 Pt 2): p. 1773-6.
213. Connolly, D.J., C.R. Lamb, and A. Boswood, *Right-to-left shunting patent ductus arteriosus with pulmonary hypertension in a cat.* J Small Anim Pract, 2003. **44**(4): p. 184-8.
214. van de Sandt, R.R., et al., *[Arterial hypertension in the cat. A pathobiologic and clinical review with emphasis on the ophthalmologic aspects].* Tijdschr Diergeneeskd, 2003. **128**(1): p. 2-10.
215. Morris, B., et al., *Polymorphism (-173G>A) in promoter of human epithelial sodium channel gamma subunit gene (SCNN1G) and association analysis in essential hypertension.* Hum Mutat, 2001. **17**(2): p. 157.
216. Wiltshire, S., et al., *Investigating the association between K198N coding polymorphism in EDN1 and hypertension, lipoprotein levels, the metabolic syndrome and cardiovascular disease.* Hum Genet, 2008. **123**(3): p. 307-13.
217. Hedegaard, C.J., et al., *Autoantibodies to myelin basic protein (MBP) in healthy individuals and in patients with multiple sclerosis: a role in regulating cytokine responses to MBP.* Immunology, 2009. **128**(1 Suppl): p. e451-61.
218. Deraos, G., et al., *Citrullination of linear and cyclic altered peptide ligands from myelin basic protein (MBP(87-99)) epitope elicits a Th1 polarized response by T cells isolated from multiple sclerosis patients: implications in triggering disease.* J Med Chem, 2008. **51**(24): p. 7834-42.
219. Stoevring, B., J.L. Frederiksen, and M. Christiansen, *CRYAB promoter polymorphisms: influence on multiple sclerosis susceptibility and clinical presentation.* Clin Chim Acta, 2007. **375**(1-2): p. 57-62.
220. Burwick, R.M., et al., *APOE epsilon variation in multiple sclerosis susceptibility and disease severity: some answers.* Neurology, 2006. **66**(9): p. 1373-83.
221. Hautecoeur, P., et al., *Variations of IL2, IL6, TNF alpha plasmatic levels in relapsing remitting multiple sclerosis.* Acta Neurol Belg, 1997. **97**(4): p. 240-3.
222. Ronaghi, M., S. Vallian, and M. Etemadifar, *CD24 gene polymorphism is associated with the disease progression and susceptibility to multiple sclerosis in the Iranian population.* Psychiatry Res, 2009. **170**(2-3): p. 271-2.
223. Otaegui, D., et al., *CD24 V/V is an allele associated with the risk of developing multiple sclerosis in the Spanish population.* Mult Scler, 2006. **12**(4): p. 511-4.
224. Babenko, S.A., et al., *[The VDR gene polymorphism in patients with multiple sclerosis].* Zh Nevrol Psikhiatr Im S S Korsakova, 2009. **109**(7 Suppl 2): p. 23-7.
225. Kantarci, O.H., et al., *A population-based study of IL4 polymorphisms in multiple sclerosis.* J Neuroimmunol, 2003. **137**(1-2): p. 134-9.
226. Lopez, E., et al., *Interferon gamma, IL2, IL4, IL10 and TNFalpha secretions in multiple sclerosis patients treated with an anti-CD4 monoclonal antibody.* Autoimmunity, 1999. **29**(2): p. 87-92.

227. Lorentzen, A.R., et al., *The SH2D2A gene and susceptibility to multiple sclerosis.* J Neuroimmunol, 2008. **197**(2): p. 152-8.
228. Dai, K.Z., et al., *The T cell regulator gene SH2D2A contributes to the genetic susceptibility of multiple sclerosis.* Genes Immun, 2001. **2**(5): p. 263-8.
229. Saarela, J., et al., *PRKCA and multiple sclerosis: association in two independent populations.* PLoS Genet, 2006. **2**(3): p. e42.
230. Barton, A., et al., *Association of protein kinase C alpha (PRKCA) gene with multiple sclerosis in a UK population.* Brain, 2004. **127**(Pt 8): p. 1717-22.
231. Palacios, R., et al., *Genomic regulation of CTLA4 and multiple sclerosis.* J Neuroimmunol, 2008. **203**(1): p. 108-15.
232. Greve, B., et al., *Multiple sclerosis and the CTLA4 autoimmunity polymorphism CT60: no association in patients from Germany, Hungary and Poland.* Mult Scler, 2008. **14**(2): p. 153-8.
233. Heidari, J., et al., *Association study of the -866G/A UCP2 gene promoter polymorphism with type 2 diabetes and obesity in a Tehran population: a case control study.* Arch Iran Med, 2010. **13**(5): p. 384-90.
234. Zuo, H.J., et al., *[Study on polymorphism of UCP2 gene in Chengdu simple obesity and normal-weight people and a preliminary investigation of its relationship with gut bacteria].* Sichuan Da Xue Xue Bao Yi Xue Ban, 2009. **40**(5): p. 865-8, 876.
235. Wang, D., et al., *Association of the MC4R V103I polymorphism with obesity: a Chinese case-control study and meta-analysis in 55,195 individuals.* Obesity (Silver Spring), 2010. **18**(3): p. 573-9.
236. Grant, S.F., et al., *Investigation of the locus near MC4R with childhood obesity in Americans of European and African ancestry.* Obesity (Silver Spring), 2009. **17**(7): p. 1461-5.
237. Johansson, L.E., et al., *Interaction between PPARG Pro12Ala and ADIPOQ G276T concerning cholesterol levels in childhood obesity.* Int J Pediatr Obes, 2009. **4**(2): p. 119-25.
238. Aller, R., et al., *Role of -55CT polymorphism of UCP3 gene on non alcoholic fatty liver disease and insulin resistance in patients with obesity.* Nutr Hosp, 2010. **25**(4): p. 572-6.
239. Jia, J.J., et al., *The polymorphisms of UCP2 and UCP3 genes associated with fat metabolism, obesity and diabetes.* Obes Rev, 2009. **10**(5): p. 519-26.
240. Sertic, J., et al., *Variants of ESR1, APOE, LPL and IL-6 loci in young healthy subjects: association with lipid status and obesity.* BMC Res Notes, 2009. **2**: p. 203.
241. Ramadori, G., et al., *SIRT1 deacetylase in POMC neurons is required for homeostatic defenses against diet-induced obesity.* Cell Metab, 2010. **12**(1): p. 78-87.
242. Parton, L.E., et al., *Glucose sensing by POMC neurons regulates glucose homeostasis and is impaired in obesity.* Nature, 2007. **449**(7159): p. 228-32.
243. Tzanavari, T., P. Giannogonas, and K.P. Karalis, *TNF-alpha and obesity.* Curr Dir Autoimmun, 2010. **11**: p. 145-56.
244. Park, E.J., et al., *Dietary and genetic obesity promote liver inflammation and tumorigenesis by enhancing IL-6 and TNF expression.* Cell, 2010. **140**(2): p. 197-208.

245. Hung, C.C., et al., *Studies of the SIM1 gene in relation to human obesity and obesity-related traits.* Int J Obes (Lond), 2007. **31**(3): p. 429-34.
246. Tolson, K.P., et al., *Postnatal Sim1 deficiency causes hyperphagic obesity and reduced Mc4r and oxytocin expression.* J Neurosci, 2010. **30**(10): p. 3803-12.
247. Benzinou, M., et al., *Common nonsynonymous variants in PCSK1 confer risk of obesity.* Nat Genet, 2008. **40**(8): p. 943-5.
248. Heni, M., et al., *Association of obesity risk SNPs in PCSK1 with insulin sensitivity and proinsulin conversion.* BMC Med Genet, 2010. **11**: p. 86.
249. Briggs, D.I., et al., *Diet-Induced Obesity Causes Ghrelin Resistance in Arcuate NPY/AgRP Neurons.* Endocrinology, 2010. **151**(10): p. 4745-55.
250. Stofkova, A., et al., *Activation of hypothalamic NPY, AgRP, MC4R, AND IL-6 mRNA levels in young Lewis rats with early-life diet-induced obesity.* Endocr Regul, 2009. **43**(3): p. 99-106.
251. Duan, C., et al., *[Molecular mechanism of SH2B1 in regulating JAK2/IRS2 during obesity development].* Zhong Nan Da Xue Xue Bao Yi Xue Ban, 2010. **35**(3): p. 209-14.
252. Feigelson, H.S., et al., *Genetic variation in candidate obesity genes ADRB2, ADRB3, GHRL, HSD11B1, IRS1, IRS2, and SHC1 and risk for breast cancer in the Cancer Prevention Study II.* Breast Cancer Res, 2008. **10**(4): p. R57.
253. Hsiao, D.J., et al., *Weight loss and body fat reduction under sibutramine therapy in obesity with the C825T polymorphism in the GNB3 gene.* Pharmacogenet Genomics, 2009. **19**(9): p. 730-3.
254. Lin, W.H., et al., *Molecular scanning of the human sorbin and SH3-domain-containing-1 (SORBS1) gene: positive association of the T228A polymorphism with obesity and type 2 diabetes.* Hum Mol Genet, 2001. **10**(17): p. 1753-60.
255. Mammes, O., et al., *Novel polymorphisms in the 5' region of the LEP gene: association with leptin levels and response to low-calorie diet in human obesity.* Diabetes, 1998. **47**(3): p. 487-9.
256. Cleveland, R.J., et al., *Common genetic variations in the LEP and LEPR genes, obesity and breast cancer incidence and survival.* Breast Cancer Res Treat, 2010. **120**(3): p. 745-52.
257. Gjesing, A.P., et al., *No consistent effect of ADRB2 haplotypes on obesity, hypertension and quantitative traits of body fatness and blood pressure among 6,514 adult Danes.* PLoS One, 2009. **4**(9): p. e7206.
258. Ben Ali, S., et al., *The G3057A LEPR polymorphism is associated with obesity in Tunisian women.* Nutr Metab Cardiovasc Dis, 2010.
259. Pyrzak, B., et al., *No association of LEPR Gln223Arg polymorphism with leptin, obesity or metabolic disturbances in children.* Eur J Med Res, 2009. **14 Suppl 4**: p. 201-4.
260. Becker, K.G., *The common variants/multiple disease hypothesis of common complex genetic disorders.* Med Hypotheses, 2004. **62**(2): p. 309-17.
261. Orr, N. and S. Chanock, *Common genetic variation and human disease.* Adv Genet, 2008. **62**: p. 1-32.
262. Zaninovich, A.A., *[Role of uncoupling proteins UCP1, UCP2 and UCP3 in energy balance, type 2 diabetes and obesity. Synergism with the thyroid].* Medicina (B Aires), 2005. **65**(2): p. 163-9.

263. Prior, S.J., A.P. Goldberg, and A.S. Ryan, *ADRB2 Haplotype Is Associated With Glucose Tolerance and Insulin Sensitivity in Obese Postmenopausal Women.* Obesity (Silver Spring), 2010.
264. Bracale, R., et al., *Metabolic syndrome and ADRB3 gene polymorphism in severely obese patients from South Italy.* Eur J Clin Nutr, 2007. **61**(10): p. 1213-9.
265. Mahajan, A., et al., *Obesity-dependent association of TNF-LTA locus with type 2 diabetes in North Indians.* J Mol Med, 2010. **88**(5): p. 515-22.
266. Mattevi, V.S., V.M. Zembrzuski, and M.H. Hutz, *Effects of a PPARG gene variant on obesity characteristics in Brazil.* Braz J Med Biol Res, 2007. **40**(7): p. 927-32.
267. Rung, J., et al., *Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia.* Nat Genet, 2009. **41**(10): p. 1110-5.
268. Ridderstrale, M., et al., *Increased risk of obesity associated with the variant allele of the PPARGC1A Gly482Ser polymorphism in physically inactive elderly men.* Diabetologia, 2006. **49**(3): p. 496-500.
269. Zhu, S., et al., *Evaluation of the association between the PPARGC1A genetic polymorphisms and type 2 diabetes in Han Chinese population.* Diabetes Res Clin Pract, 2009. **86**(3): p. 168-72.
270. Bougneres, P., *Genetics of obesity and type 2 diabetes: tracking pathogenic traits during the predisease period.* Diabetes, 2002. **51 Suppl 3**: p. S295-303.
271. Vimaleswaran, K.S. and R.J. Loos, *Progress in the genetics of common obesity and type 2 diabetes.* Expert Rev Mol Med, 2010. **12**: p. e7.
272. Qi, L., et al., *Common variations in perilipin gene, central obesity, and risk of type 2 diabetes in US women.* Obesity (Silver Spring), 2008. **16**(5): p. 1061-5.
273. Elbers, C.C., et al., *A strategy to search for common obesity and type 2 diabetes genes.* Trends Endocrinol Metab, 2007. **18**(1): p. 19-26.
274. Raz, I., *Complex impact of obesity on type 2 diabetes.* Isr Med Assoc J, 2005. **7**(6): p. 402-3.
275. Lou, Y., et al., *A46G and C79G polymorphisms in the beta2-adrenergic receptor gene (ADRB2) and essential hypertension risk: a meta-analysis.* Hypertens Res, 2010.
276. Lee, S.J., W.J. Kim, and S.K. Moon, *TNF-alpha regulates vascular smooth muscle cell responses in genetic hypertension.* Int Immunopharmacol, 2009. **9**(7-8): p. 837-43.
277. Yeo, E.S., et al., *Tumor necrosis factor (TNF-alpha) and C-reactive protein (CRP) are positively associated with the risk of chronic kidney disease in patients with type 2 diabetes.* Yonsei Med J, 2010. **51**(4): p. 519-25.
278. Liu, L., et al., *Pro12Ala polymorphism in the PPARG gene contributes to the development of diabetic nephropathy in Chinese type 2 diabetic patients.* Diabetes Care, 2010. **33**(1): p. 144-9.
279. Gundogan, K., et al., *Prevalence of metabolic syndrome in the Mediterranean region of Turkey: evaluation of hypertension, diabetes mellitus, obesity, and dyslipidemia.* Metab Syndr Relat Disord, 2009. **7**(5): p. 427-34.
280. Kristiansen, O.P., et al., *Association of a functional 17beta-estradiol sensitive IL6-174G/C promoter polymorphism with early-onset type 1 diabetes in females.* Hum Mol Genet, 2003. **12**(10): p. 1101-10.

281. Johansen, A., et al., *IRS1, KCNJ11, PPARgamma2 and HNF-1alpha: do amino acid polymorphisms in these candidate genes support a shared aetiology between type 1 and type 2 diabetes?* Diabetes Obes Metab, 2006. **8**(1): p. 75-82.
282. Eftychi, C., et al., *Analysis of the type 2 diabetes-associated single nucleotide polymorphisms in the genes IRS1, KCNJ11, and PPARG2 in type 1 diabetes.* Diabetes, 2004. **53**(3): p. 870-3.
283. Ponting, C.P. and L. Goodstadt, *Statistical genetics: usual suspects in complex disease.* Eur J Hum Genet, 2005. **13**(3): p. 269-70.
284. Botstein, D. and N. Risch, *Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.* Nat Genet, 2003. **33 Suppl**: p. 228-37.