ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΜΕΤΑ-ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ ΓΙΑ ΤΑ ΚΑΡΔΙΑΓΓΕΙΑΚΑ ΝΟΣΗΜΑΤΑ

Μπογιατζή Σπυριδούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Μπάγκος Παντελής

Αναπληρωτής Καθηγητής

ΛΑΜΙΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2016

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΜΕΤΑ-ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ ΓΙΑ ΤΑ ΚΑΡΔΙΑΓΓΕΙΑΚΑ ΝΟΣΗΜΑΤΑ

Μπογιατζή Σπυριδούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Μπάγκος Παντελής

Αδάμ Μαρία

Πλαγιανάκος Βασίλειος

ΛΑΜΙΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2016

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον Αναπληρωτή Καθηγητή κ. Παντελή Μπάγκο για την δυνατότητα που μου έδωσε να πραγματοποιήσω την πτυχιακή μου εργασία και την εμπιστοσύνη που μου έδειξε. Οι σημαντικές υποδείξεις και συμβουλές του με κατεύθυναν σ' ένα σωστό τρόπο σκέψης και μου προσέφεραν σημαντικά εφόδια.

Επίσης θα ήθελα να ευχαριστήσω εκ' βαθέων την υποψήφια Διδάκτορα Παναγιώτα Κοντού για το πολύτιμο χρόνο που διέθεσε για την περάτωση της παρούσας εργασίας.

Τέλος, θέλω να εκφράσω ένα τεράστιο ευχαριστώ στην οικογένεια μου, για την στήριξη και την εμπιστοσύνη που μου έδειξε όλα αυτά τα χρόνια των σπουδών μου.

Περιεχόμενα

Περίληψη

Το έμφραγμα του μυοκαρδίου είναι από τις κυριότερες αιτίες θανάτου στις ανεπτυγμένες χώρες. Κύριοι παράγοντες που προκαλούν την ασθένεια είναι το κάπνισμα, η παχυσαρκία, η χοληστερίνη, το στρες αλλά και μη τροποποιήσιμοι παράγοντες όπως η ηλικία, το φύλο και η κληρονομικότητα. Σε αυτή την πτυχιακή εργασία έγινε προσπάθεια να εντοπιστούν οι γονιδιακοί παράγοντες που οδηγούν στο έμφραγμα του μυοκαρδίου μέσω της μετα-ανάλυσης μικροσυστοιχιών, ώστε να προβλεφθεί ο σχετικός κίνδυνος εμφάνισης της ασθένειας από ένα υγιές άτομο. Η μετα-ανάλυση των μικροσυστοιχιών, είναι μια μέθοδος στατιστικής ανάλυσης. Μελετώντας ανεξάρτητες έρευνες για την ίδια γενική υπόθεση, αποκτώνται πιο αντιπροσωπευτικές και ακριβείς εκτιμήσεις της διαφορικής έκφρασης γονιδίων.

Από τη βιβλιογραφική αναζήτηση στις βάσεις δεδομένων Pubmed και GEO, συγκεντρώθηκαν μελέτες μικροσυστοιχιών γονιδιακής έκφρασης με δείγματα υγιών-ασθενών. Κατεγράφησαν τα χαρακτηριστικά της κάθε μελέτης όπως αριθμός συμμετεχόντων και πλατφόρμα μικροσυστοιχίας. Αφού σε κάθε μελέτη έγινε ο αναγκαίος στατιστικός έλεγχος, μέσω της μεθόδου bootstrap με αναδειγματοληψία και επανάθεση, έπειτα εκτελέστηκε η μετα-ανάλυση. Ως μέγεθος επίδρασης επιλέχτηκε η διαφορά μέσων τιμών, μέσω του ελέγχου t και εφαρμόστηκε random effects model. Τέλος, εφαρμόστηκαν μέθοδοι διόρθωσης του p-value και με τη βοήθεια του Biocompendium αποκαλύφθηκαν τα βιοχημικά μονοπάτια, η μοριακή λειτουργία και το δίκτυο αλληλεπιδράσεων μεταξύ των στατιστικά σημαντικών γονιδίων.

Χρησιμοποιήθηκαν 4 μελέτες μικροσυστοιχιών από τη βάση δεδομένων GEO με 93 ασθενείς και 89 υγιείς. Η μετα-ανάλυση πραγματοποιήθηκε σε 31.180 γονίδια με τη χρήση του στατιστικού πακέτου STATA. Από τα αποτελέσματα της μετα-ανάλυσης εντοπίστηκαν 666 στατιστικά σημαντικά γονίδια κατά FDR. Από το bioCompendium βρέθηκαν 118 βιοχημικά μονοπάτια κάποια από τα οποία είναι η αλληλεπίδραση των κυτταροκινών με τους υποδοχείς τους, ο τύπου 1 σακχαρώδης διαβήτης, η νόσος του Πάρκινσον, ο καρκίνος στο πάγκρεας και η ιογενής μυοκαρδίτιδα.

Από τη μετα-ανάλυση βρέθηκαν γονίδια τα οποία εμφανίζουν στατιστικά σημαντική συσχέτιση με το έμφραγμα του μυοκαρδίου και οι επιμέρους μελέτες δε

μπορούν να τα ανιχνεύσουν. Ωστόσο χρειάζεται περαιτέρω μελέτη, για να αξιολογηθεί η σχέση όλων αυτών των γονιδίων με την ασθένεια του μυοκαρδίου.

Abstract

Myocardial infarction is one of the major causes of death in developed countries. The main factors that cause the disease are smoking, obesity, cholesterol, stress and non-modifiable factors such as age, gender and heredity. In this thesis there was an attempt to identify the genomic factors leading to myocardial infarction through post-microarray analysis to predict the relative risk of disease of that of a healthy person. The meta-analysis of microarray, is a statistical analysis method. Studying independent studies on the same general assumption, acquired more representative and accurate estimates of differential gene expression.

From a literature search on Pubmed and GEO databases microarray gene expression studies with healthy-patient samples were collected. They recorded the characteristics of each study such as number of participants, and microarray platform.In each study was the necessary statistical check through bootstrap resampling method, resetting and then performed a meta-analysis. As an effect size averaging difference was selected by controlling t and applied random effect model. Finally applied were the correction methods of p-value with the aid of the disclosed Biocompendium biochemical pathways, molecular function and the network of interactions between the statistically significant genes.

Four studies used microarrays from GEO database of 93 patients of which 89 were healthy. The meta-analysis was performed on 31,180 genes using the STATA statistical package. The results of this meta-analysis identified 666 genes significantly in FDR. By using BioCompendium platform, 118 biochemical pathways appear to majority of genes belonging to the interaction of cytokines with their receptors, type 1 diabetes mellitus, Parkinson's disease, cancer of the pancreas, and viral myocarditis.

The results obtained from the meta-analysis have a statistically significant correlation with myocardial infarction that individual studies cannot detect. However needs further exploration, to evaluate the relationship of all of these genes in myocardial disease.

1. ΕΙΣΑΓΩΓΗ

1.1. ΚΑΡΔΙΑΓΓΕΙΑΚΕΣ ΠΑΘΗΣΕΙΣ

1.1.1. Είδη Καρδιαγγειακών Παθήσεων

Οι καρδιαγγειακές ασθένειες αποτελούν μια ομάδα διαταραχών της καρδιάς και των αιμοφόρων αγγείων. Οι καρδιοπάθειες διακρίνονται σε συγγενείς και επίκτητες. Μπορούν να ομαδοποιηθούν σε τρεις τύπους: τις στεφανιαίες νόσους, τις διαταραχές των καρδιακών βαλβίδων και την καρδιομυοπάθεια.

Στεφανιαία νόσος. Στένωση των αιμοφόρων αγγείων που αιματώνουν την καρδιά.

Αγγειακό εγκεφαλικό επεισόδιο. Διαταραχή των αιμοφόρων αγγείων που αιματώνουν τον εγκέφαλο.(ισχαιμικό επεισόδιο-αιμορραγικό επεισόδιο)

Περιφερική αρτηριοπάθεια. Νόσος των περιφερικών αγγείων που τροφοδοτούν τα άνω και κάτω άκρα.

Ρευματική καρδιοπάθεια. Καταστροφή του καρδιακού μυ και των βαλβίδων της καρδιάς από ρευματικό πυρετό, που προκαλείται από λοίμωξη στρεπτόκοκκου.

Συγγενής καρδιοπάθεια. Ανωμαλίες των δομών της καρδιάς οι οποίες υπάρχουν από τη γέννηση.

Εν τω Βάθει Φλεβοθρόμβωση και πνευμονική εμβολή. Απόφραξη των αγγείων των κάτω άκρων με θρόμβους, οι οποίοι μπορούν να μεταφερθούν στην καρδιά και στον πνεύμονα.

Ανευρύσματα και διαχωρισμός αορτής. Πρόκειται για διάταξη και ρήξη της αορτής.

Άλλα καρδιαγγειακά νοσήματα. Αρτηριακή υπέρταση, όγκοι καρδιάς, εγκεφαλικά ανευρύσματα, καρδιομυοπάθεια, βαλβιδοπάθειες(ΣΙΜΟΥ, 2008).

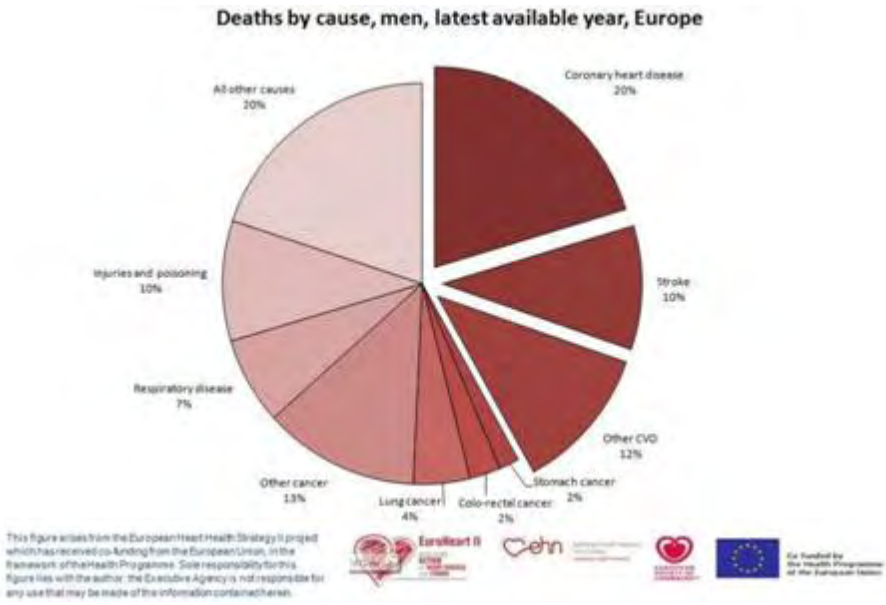
1.1.2. Στατιστικές Αναφορές

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, τα καρδιαγγειακά νοσήματα είναι η νούμερο ένα αιτία θνησιμότητας και νοσηρότητας παγκοσμίως. Υπολογίζεται ότι 17,5 εκατομμύρια άνθρωποι έχασαν τη ζωή τους το 2012 από καρδιαγγειακά αντιπροσωπεύοντας το 31% όλων των θανάτων παγκοσμίως. Η πλειοψηφία των θανάτων αυτών σημειώνεται στις αναπτυσσόμενες χώρες. Στην εικόνα 1.5 παρουσιάζεται ο δείκτης θνησιμότητας από καρδιαγγειακά(Mendis, 2014).

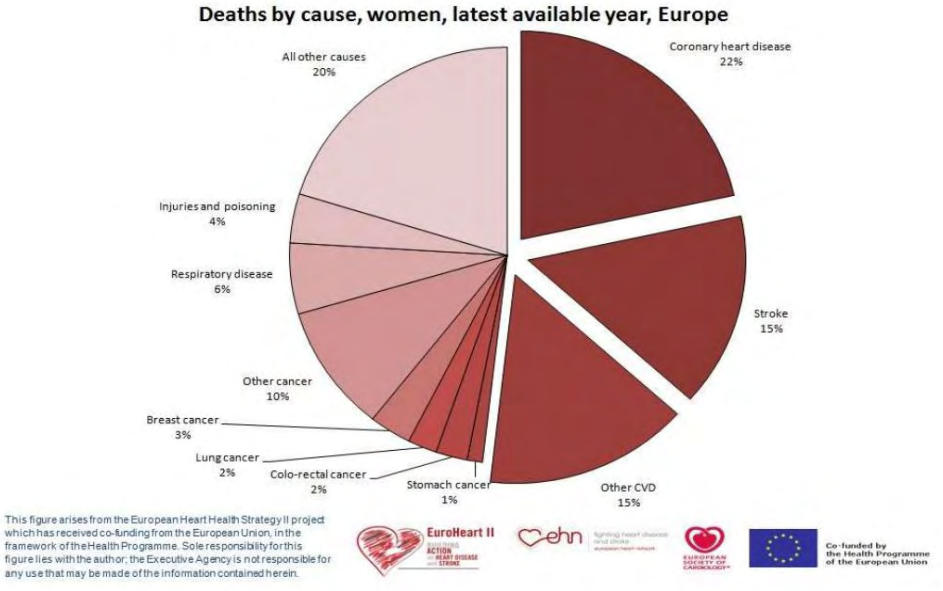
Ταυτόχρονα η Ευρωπαϊκή Καρδιολογική Κοινότητα δηλώνει ότι κάθε χρόνο τα καρδιαγγειακά προκαλούν πάνω από 4 εκατομμύρια θανάτους στην Ευρώπη. Αποτελούν το 47% όλων των θανάτων στην Ευρώπη και είναι η κύρια αιτία θανάτου στις γυναίκες. Ευθύνονται για περισσότερους θανάτους από ότι όλοι οι καρκίνοι μαζί. Εκτιμάται ότι κοστίζουν στην Ευρωπαϊκή οικονομία 196 δισεκατομμύρια ευρώ ετησίως. Στην εικόνα 1.1 και 1.2 παρουσιάζονται οι κυριότερες αιτίες θανάτου ανδρών και γυναικών.

Κύριοι παράγοντες κινδύνου για αυτές τις ασθένειες είναι το κάπνισμα, το παθητικό κάπνισμα, η παχυσαρκία, η καθιστική ζωή, η δυσλιπιδαιμία, η αρτηριακή υπέρταση, ο σακχαρώδης διαβήτης, κληρονομικό ιστορικό, ψυχολογικό στρες, υπερκατανάλωση αλκοόλ, ηλικία, φύλο, κοινωνικοοικονομικό επίπεδο. Όσοι έχουν περισσότερους από έναν παράγοντες κινδύνου εμφανίζουν και μεγαλύτερο κίνδυνο να εμφανίσουν κάποιο καρδιαγγειακό νόσημα(Løgstrup & O'Kelly, 2012).

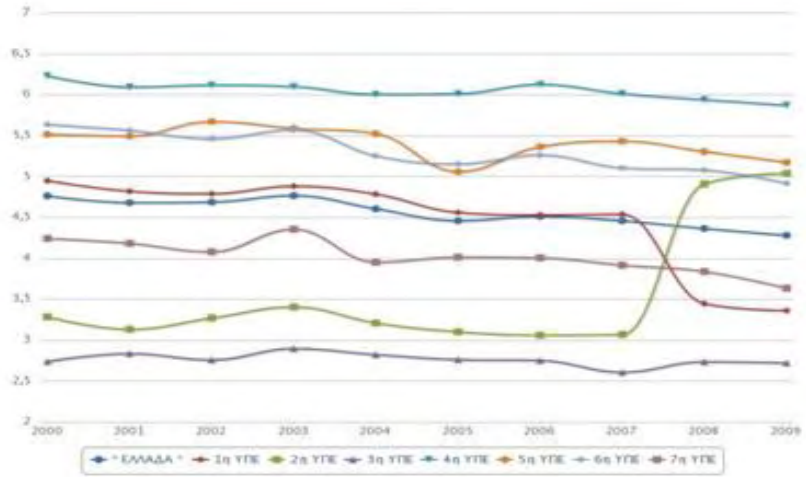
Συμπερασματικά τα καρδιαγγειακά είναι νοσήματα στα οποία κοινός παρανομαστής είναι η παρουσία αθηροσκλήρωσης στα αγγεία του σώματος. Εμφανίζονται αιφνίδια και οφείλονται είτε σε κληρονομικούς παράγοντες είτε σε επίκτητους. Αφορούν όλο τον πληθυσμό και ιδιαίτερα τις γυναίκες και τις χαμηλότερες κοινωνικοοικονομικές ομάδες. Δεν είναι λοιπόν απαραίτητη μόνο η μείωση των παραγόντων κινδύνου αλλά και η βελτίωση της πρώιμης διάγνωσης των νοσημάτων αυτών, από τον επιστημονικό και ιατρικό κλάδο.



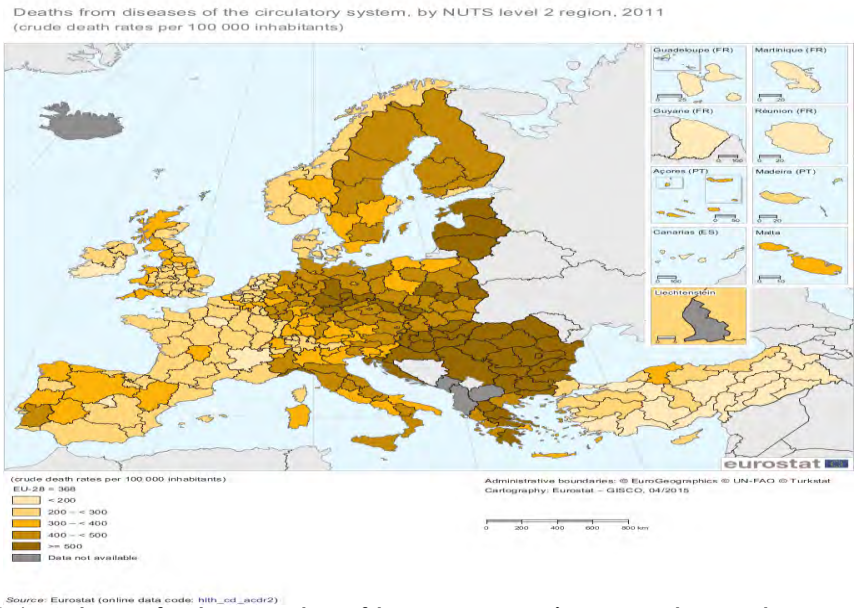
Εικόνα 1.1 Αιτίες θανάτου ανδρών στην Ευρώπη. Πηγή : European society of cardiology



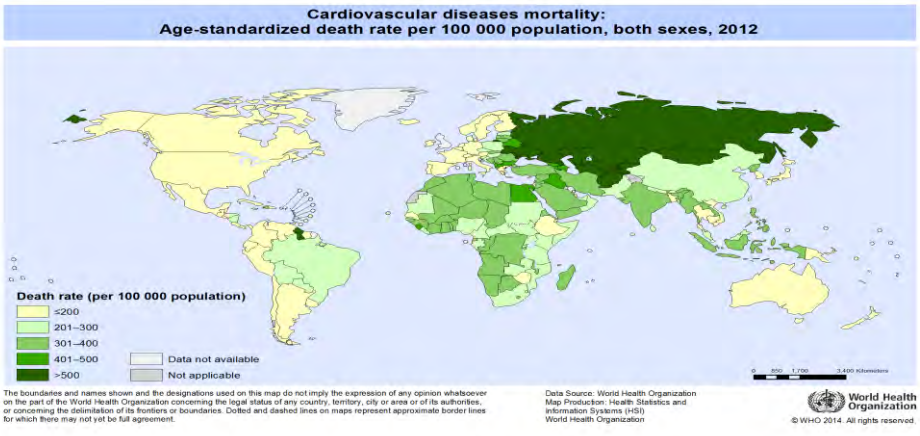
Εικόνα1.2 Αιτίες θανάτου γυναικών στην Ευρώπη. Πηγή: European society of cardiology



Εικόνα 1.3 Δείκτης θανάτων από καρδιαγγειακές παθήσεις για άντρες-γυναίκες, σε όλες τις Υγειονομικές Περιοχές της Ελλάδας(2000-2009). Πηγή: Υγειονομικός χάρτης



Εικόνα 1.4 Δείκτης θανάτων από ασθένειες του κυκλοφοριακού συστήματος ανά 100.000 κατοίκους(2011). Πηγή Eurostat

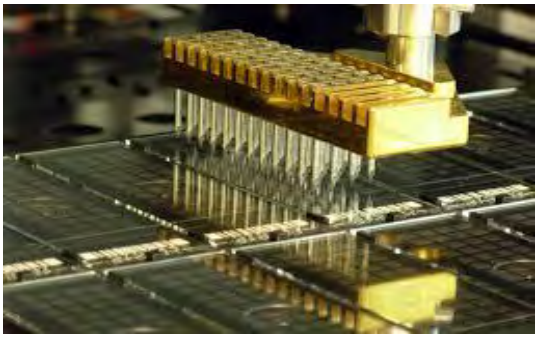


Εικόνα 1.5 Δείκτης θανάτων από καρδιαγγειακές ασθένειες και για τα δύο φύλλα ανά 100.000 κατοίκους(2012). Πηγή WHO

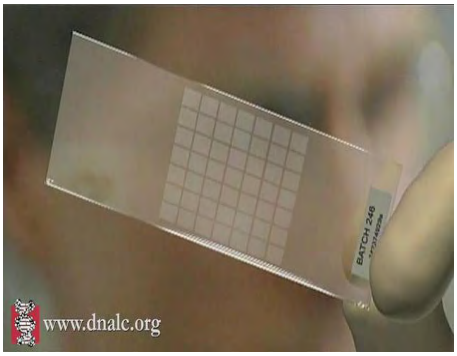
1.2. ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ

1.2.1. Κατασκευή

Οι μικροσυστοιχίες DNA είναι ένα εργαλείο, που από το 90’ και έπειτα βοηθά τους επιστήμονες να αναλύσουν την γονιδιακή έκφραση με ταυτόχρονη μελέτη μεγάλου αριθμού γονιδίων. Γονιδιακή έκφραση είναι η μεταγραφή του DNA σε RNA και η μετάφρασή του σε πρωτεΐνες. Ουσιαστικά είναι ένα πολύ μικρό γυάλινο πλακίδιο με 20000 κηλίδες(spot) στις οποίες ακινητοποιούνται ανιχνευτές(probes). Κάθε spot έχει μέγεθος 5-150 μm, και χάρη στο ρομποτικό βραχίονα όλα τα spot έχουν το ίδιο σχήμα και ισαπέχουν μεταξύ τους. Τα probes είναι κομμάτια DNA, cDNA ή ολιγονουκλεοτίδια που αντιπροσωπεύουν ένα γνωστό γονίδιο. Συνεπώς η μικροσυστοιχία έχει 20000 spot και σε κάθε ένα συγκρατείται ένα γονίδιο. Στις εικόνες 1.6 και 1.7 φαίνονται ο ρομποτικός βραχίονας και το πλακίδιο μικροσυστοιχίας (Bumgarner, 2013).



Εικόνα 1.6 Ρομποτικός βραχίονας



Εικόνα 1.7 Πλακίδιο μικροσυστοιχίας

Οι πιο γνωστές πλατφόρμες μικροσυστοιχιών είναι της Affymetrix, της Illumina και της Agilent. Τα πειράματα σε μικροσυστοιχίες DNA χωρίζονται σε τρεις κατηγορίες ανάλογα με το είδος του δείγματος που τοποθετείται στα probes και τα αποτελέσματα που προσκομίζονται(Ανάλυση έκφρασης μικροσυστοιχιών (cDNA), Μικροσυστοιχίες για ανάλυση μεταλλάξεων(gDNA),Υβριδοποίηση συγκριτικής γονιδιωματικής).

Μερικές από τις εφαρμογές των μικροσυστοιχιών είναι η ανακάλυψη λειτουργιών γονιδίων, η εύρεση του τρόπου συντονισμού γονιδίων μεταξύ τους, η δημιουργία νέων φαρμάκων εξειδικευμένων για μια ασθένεια, η καλύτερη κατανόηση μιας ασθένειας, ο έλεγχος απόκρισης σε θεραπείες και η σύγκριση επιπέδων έκφρασης γονιδίων. Μπορούν να χρησιμοποιηθούν για την ανάλυση της γονιδιακής έκφρασης σε ένα δείγμα ή για σύγκριση σε δυο διαφορετικούς κυτταρικούς τύπους ή σε υγιή - ασθενή ιστό.

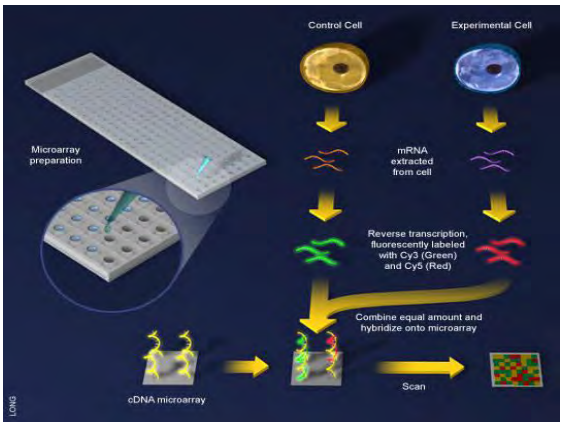
1.2.2. Πειραματική Διαδικασία

Το δεύτερο βήμα, μετά τη δημιουργία και επιλογή της μικροσυστοιχίας, είναι η απομόνωση του mRNA από τον υπό μελέτη ιστό ή κύτταρα. Για τη μελέτη μιας ασθένειας, χρειάζεται δείγμα ιστού από υγιή και ασθενή κύτταρα. Αφού μετατραπεί το mRNA σε cDNA(με την αντίστροφη μεταγραφάση)σημαίνεται με φθορίζουσες χρωστικές. Συνηθίζεται με πράσινη χρωστική να σημαίνονται τα υγιή κύτταρα και με κόκκινη τα ασθενή.

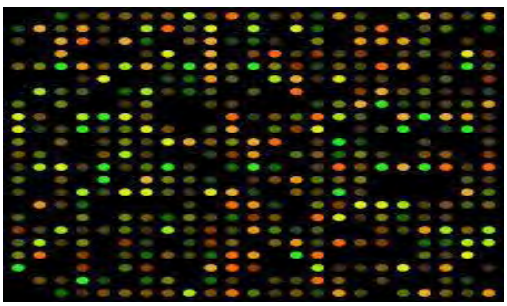
Επόμενο βήμα είναι ο υβριδισμός. Όταν δύο συμπληρωματικές μονόκλωνες αλυσίδες DNA βρεθούν στον ίδιο χώρο κάτω από κατάλληλες συνθήκες, τότε υβριδοποιούνται και δημιουργούν δίκλωνο μόριο. Τοποθετούνται λοιπόν τα ιχνηθετημένα μόρια cDNA υγιών και ασθενών στη μικροσυστοιχία και επωάζονται στους 42°C. Μετά από ώρα κάποια από τα μόρια θα υβριδοποιηθούν με τα μόριαDNA(ανιχνευτές) των spot και κάποια όχι. Η μικροσυστοιχία πλένεται ώστε να αφαιρεθούν τα μόρια που δεν υβριδοποιήθηκαν.

Οι υβριδισμένες επιφάνειες μικροσυστοιχιών σαρώνονται με τη βοήθεια ενός σαρωτή που μετατρέπει την ένταση σήματος σε ψηφιακή εικόνα. Αναλύει τις εικόνες των μικροσυστοιχιών, διαχειρίζεται τα δεδομένα και διεξάγει τη βασική ανάλυση της έκφρασης των γονιδίων. Πρώτα θα σαρώσει το πλακίδιο στο μήκος κύματος της πράσινης χρωστικής και έπειτα στο μήκος κύματος της κόκκινης χρωστικής. Δημιουργούνται έτσι ψηφιακές εικόνες που φανερώνουν την ένταση φθορισμού για

κάθε κηλίδα. Ο συνδυασμός των δύο εικόνων διευκολύνει την οπτική αναγνώριση των υπέρ εκφρασμένων γονιδίων. Επειδή ο υβριδισμός γίνεται ταυτόχρονα και για τα δύο δείγματα θα υπάρχει ανταγωνισμός μεταξύ των στόχων και των ανιχνευτών. Μια μεγάλη ένταση φθορισμού υποδηλώνει πως το γονίδιο σε αυτή την κηλίδα(spot) ήταν πολύ ενεργό(παρήγαγε πολλά μόρια m RNA, που υβριδοποιήθηκαν με το cDNA) ενώ μια χαμηλή ένταση φθορισμού δείχνει λιγότερο ενεργά γονίδια. Οι πράσινες και κόκκινες κηλίδες φανερώνουν ποια γονίδια υβριδοποιήθηκαν και βρέθηκαν σε αφθονία στον υγιή και ασθενή ιστό αντίστοιχα. Τα κίτρινα σημεία δείχνουν πως η έκφραση του γονιδίου ήταν παρόμοια και στον υγιή και στον ασθενή ιστό, ενώ οι μαύρες κηλίδες δείχνουν ότι δεν έγινε υβριδισμός και άρα τα γονίδια σε αυτά τα σημεία είναι ανενεργά. Αυτό είναι λογικό αφού όλα τα κύτταρα έχουν το ίδιο γονιδίωμα αλλά εκφράζουν διαφορετικά γονίδια. Για να ποσοτικοποιηθούν τα αποτελέσματα της εικόνας και να αφαιρεθούν πιθανά σφάλματα χρειάζεται περαιτέρω ανάλυση (Trevino V, 2007).



Εικόνα 1.8 Δείγμα από φυσιολογικό κύτταρο και ασθενές απομονώνεται, φθορίζεται και υβριδοποιείται.



Εικόνα 1.9 Εικόνα από πλακίδιο μικροσυστοιχίας μετά από σάρωση

1.2.3. Κανονικοποίηση

Για να επαληθευτούν τα αποτελέσματα των μικροσυστοιχιών και να ελαχιστοποιηθούν τα σφάλματα γονιδιακής έκφρασης χρησιμοποιείται μια μαθηματική διαδικασία η κανονικοποίηση. Η κανονικοποίηση μετατρέπει τις τιμές έντασης των μικροσυστοιχιών σε συγκρίσιμες τιμές. Σφάλματα της έντασης φθορισμού μπορεί να προκύψουν από τη σήμανση του cDNA, από τις χημικές ιδιότητες χρωστικών που χρησιμοποιούνται, από την ομοιογένεια της υβριδοποίησης και τις συνθήκες σάρωσης των πλακιδίων. Η κανονικοποίηση μπορεί να γίνει μέσα στο ίδιο πλακίδιο, σε ζευγάρια πλακιδίων ή μεταξύ πολλαπλών πλακιδίων. Χρησιμοποιείται το σύνολο των γονιδίων προς κανονικοποίηση ή ένα υποσύνολο γονιδίων σταθερής έκφρασης. Τα είδη της κανονικοποίησης είναι η ολική κανονικοποίηση, η μέση λογαριθμική κεντροποίηση και κανονικοποίηση lowess.

Κανονικοποίηση ολικής έντασης: Υποθέτει ότι τα δείγματα ελέγχου και αναφοράς που συγκρίνονται έχουν την ίδια αρχική ποσότητα mRNA και ότι οι ανιχνευτές της μικροσυστοιχίας αναπαριστούν τυχαία δειγματοληψία των γονιδίων, ώστε η ένταση υβριδισμού τους να είναι ίση στο κάθε δείγμα. Υπολογίζεται παράγοντας κανονικοποίησης όπου Rk και Gk είναι οι τιμές έκφρασης του γονιδίου k στο δείγμα ελέγχου και αναφοράς και Narray είναι ο συνολικός αριθμός των υπό μελέτη γονιδίων $N_{total} = \frac{\sum_{g=1}^{N_{array}} R_k}{\sum_{g=1}^{N_{array}} G_k}$.

Ορίζονται $R'_k = R_k$ και $G'_k = N_{total} * G_k$ και ο κανονικοποιημένος λόγος για κάθε γονίδιο θα είναι: $T'_k = \frac{R'_k}{G'_k} = \frac{1}{N_{total}} * \frac{R_k}{G_k}$

μέση λογαριθμική κεντροποίηση: Υποθέτει ότι ο μέσος όρος του $\log_2(\text{ratio})$ θα πρέπει να είναι για όλα τα γονίδια μηδέν. Είναι ευαίσθητη μέθοδος για γονίδια με ακραίες τιμές έκφρασης. Παράγοντας κανονικοποίησης: $N_{mic} = \frac{\sum_{k=1}^{N_{array}} \log_2\left(\frac{R_k}{G_k}\right)}{N_{array}}$

Ορίζονται $\mathbf{R}'_k = \mathbf{R}_k$ και $\mathbf{G}'_k = 2^{N_{mic}} * \mathbf{G}_k$ και ο κανονικοποιημένος λόγος γίνεται:

$$T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{G_k * 2^{N_{mic}}} = \frac{T_k}{2^{N_{mic}}}$$

κανονικοποίηση lowess: Απομακρύνει λάθη έντασης από τους λόγους $\log_2$(R/G). Δημιουργείται διάγραμμα με άξονες το $\log_2$(Ri/Gi) και το $\log_{10}$(Ri*Gi). Η μέθοδος εντοπίζει τις αποκλίσεις και τις διορθώνει εκτελώντας μια τοπικά σταθμισμένη γραμμική παλινδρόμηση για κάθε σημείο στην καμπύλη. Αφαιρεί την υπολογισμένα βέλτιστη προσαρμογή του μέσου $\log_2$(R/G) από την πειραματικά παρατηρούμενη τιμή του λογαριθμικού λόγου για κάθε σημείο σαν συνάρτηση της $\log_{10}$(Ri*Gi) (Yang, Yee Hwa et al., 2002).

1.2.4. Στατιστική Ανάλυση

Αφότου ολοκληρωθούν τα βιολογικά πειράματα μικροσυστοιχιών αποκτώνται τα αποτελέσματά τους, δηλαδή οι τιμές έκφρασης όλων των γονιδίων. Για να μελετηθούν τα αποτελέσματα αυτά και να βρεθεί το πώς σχετίζονται μεταξύ τους, ποια η κοινή τους λειτουργία και πόσο στατιστικά σημαντικά είναι εφαρμόζονται τυποποιημένες στατιστικές προσεγγίσεις όπως το t-test, η αλλαγή διπλώματος(fold change), η permute μέθοδος και η bootstrap μέθοδος.

Η **αλλαγή διπλώματος** είναι ένας αριθμός που περιγράφει το μέτρο της διαφοράς αλλαγής από μια αρχική τιμή σε μια τελική. Προσδίδει διαφορετική τιμή σε κάθε γονίδιο και αναπαριστά το κατά πόσες φορές έχει αλλάξει η έκφραση του γονιδίου στο πειραματικό δείγμα από το βιολογικό δείγμα. Αν η τιμή είναι θετική υπάρχει υπέρ έκφραση γονιδίου, αν είναι αρνητική υπάρχει υπό έκφραση.

Το **t-test** είναι ένας στατιστικός έλεγχος που χρησιμοποιεί δεδομένα από ένα δείγμα για να ελέγξει υποθέσεις που σχετίζονται με τη μέση τιμή ενός πληθυσμού όταν η διακύμανση του αρχικού πληθυσμού είναι άγνωστη. Γίνεται εκτίμηση της διακύμανσης του πληθυσμού με τη βοήθεια της διακύμανσης του δείγματος. Ελέγχει κατά πόσο η διαφορά μέσων τιμών ανάμεσα σε δύο πληθυσμούς μπορεί να προέκυψε από τύχη κατά την επιλογή του δείγματος. Για τον τύπο του independent t-test ισχύει

$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$ (όπου $\bar{x}_1 - \bar{x}_2$ = διαφορά μέσων τιμών δείγματος και $\mu_1 - \mu_2$ = διαφορά μέσων τιμών πληθυσμού). $S_{\bar{x}_1 - \bar{x}_2}$ είναι το τυπικό σφάλμα (standard error) και ισούται με $\sqrt{\frac{S^2_{pooled}}{n_1} + \frac{S^2_{pooled}}{n_2}}$ όπου $S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$. Το μέτρο ελέγχου των αποτελεσμάτων t-test είναι το **p-value**. Σε μια στατιστική ανάλυση υπάρχει μια μηδενική υπόθεση H_0 και μια εναλλακτική υπόθεση H_1. Η μηδενική υπόθεση συνήθως αναφέρει ότι δεν υπάρχει σημαντική διαφορά μεταξύ των δύο ομάδων που μελετώνται, ενώ η εναλλακτική υπόθεση αναφέρει ότι υπάρχει διαφορά. Η τιμή p-value είναι η πιθανότητα η στατιστική συνάρτηση (π.χ t-test) να λάβει μια ακραία τιμή όταν η H_0 υπόθεση είναι αληθινή. Είναι το μικρότερο επίπεδο σημαντικότητας(α) στο οποίο η μηδενική υπόθεση μπορεί να απορριφθεί. Το επίπεδο σημαντικότητας είναι συνήθως 5%(α=0,05) ή 1%(α=0,01). Όσο πιο μικρή τιμή έχει το p-value, τόσο μεγαλύτερη είναι η βεβαιότητα ότι απορρίπτεται η H_0. Η τιμή του p κυμαίνεται ανάμεσα στο 0 και στο 1, και κάθε γονίδιο εμφανίζει μια τιμή p-value που προσδιορίζει το κατά πόσο εμφανίστηκε τυχαία ως διαφορικά εκφρασμένο. Όμως ο μεγάλος αριθμός γονιδίων που μελετώνται στις μικροσυστοιχίες εμφανίζει και πολλά εσφαλμένα αποτελέσματα ως σημαντικά (false positive). Οι μέθοδοι διόρθωσης του p-value που περιορίζουν αυτά τα σφάλματα είναι οι FDR (Simes) – Bonferroni -Holm - Sidak – Holland (Fay & Gerow, 2013).

Στις μελέτες μικροσυστοιχίας είναι σύνηθες να υπάρχει μικρό μέγεθος δείγματος και μη κανονική κατανομή των τιμών έκφρασης. Συνηθίζεται λοιπόν ο στατιστικός έλεγχος t να επακολουθείτε από τη μέθοδο permute ή τη bootstrap, ώστε να παραχθεί εκ νέου τυπικό σφάλμα και πιο αξιόπιστα διαστήματα εμπιστοσύνης. Η μέθοδος **permutation** είναι ένας έλεγχος στατιστικής σημαντικότητας, όπου υπολογίζει τη δειγματική κατανομή κάθε στατιστικού ελέγχου. Υπολογίζει όλες τις πιθανές τιμές του στατιστικού ελέγχου με αναδιατάξεις των ετικετών(label) στα παρατηρούμενα δεδομένα. Κάθε μετάθεση των ετικετών για τα δείγματα (case-control) θεωρείται πιθανή και αντιπροσωπεύει ένα τυχαίο περιστατικό των δεδομένων, υπό τη μηδενική υπόθεση της μη σύνδεσης. Λόγω του μεγάλου αριθμού των πιθανών μεταθέσεων, επιλέγεται ένα τυχαίο δείγμα μεταθέσεων για να δημιουργηθεί διαδοχικά ένα τυχαίο

δείγμα με βάση το προηγούμενο τυχαίο δείγμα. Σε κάθε τυχαίο δείγμα υπολογίζεται ο στατιστικός έλεγχος που επιθυμείται. Το επίπεδο σημαντικότητας στη συνέχεια υπολογίζεται ως η αναλογία των δειγμάτων τυχαιοποίησης με ένα στατιστικό τεστ. Η μέθοδος permutation υπερτερεί στο ότι μπορεί να χρησιμοποιηθεί ακόμα και αν δεν είναι γνωστή η κατανομή, όμως απαιτεί πολλές επαναλήψεις και δεν δουλεύει καλά για ζευγαρωτά δεδομένα. Ο κώδικας που εκτελείται είναι ο εξής

```
permute type t=r(t), reps(1000):   ttest x,by(type) uneq
```

Η **Bootstrap** είναι μια στατιστική τεχνική που εισήγαγε πρώτος ο Bradley Efron το 1979. Βασίζεται στη μέθοδο της αναδειγματοληψίας και χρησιμοποιείται για τον υπολογισμό της δειγματικής κατανομής στατιστικών χωρίς τη χρήση της κανονικής θεωρίας(π.χ ttest). Ως μέθοδος επαναδειγματοληψίας δίνει πιο ακριβείς απαντήσεις για το τυπικό σφάλμα και για τα διαστήματα εμπιστοσύνης. Στα πειράματα μετα-ανάλυσης μικροσυστοιχιών είναι αρκετά αποτελεσματική όταν η κατανομή μιας στατιστικής είναι άγνωστη και το μέγεθος του δείγματος είναι μικρό. Η μέθοδος ακολουθεί τρία βήματα, αναδειγματοληψία, υπολογισμό της Bootstrap κατανομής και χρήση της κατανομής για εξαγωγή συμπερασμάτων. Έστω ότι υπάρχει ένα σύνολο δεδομένων N. Αρχικά, κατά τον αλγόριθμο monte carlo δημιουργούνται χιλιάδες καινούργια δείγματα ίδιου μεγέθους με το αρχικό σύνολο δεδομένων έπειτα από τυχαία αναδειγματοληψία και επανατοποθέτηση από το αρχικό πραγματικό δείγμα. Εναλλακτικά πρέπει να υπολογιστούν N^N νέα πιθανά δείγματα. Λόγω της επανατοποθέτησης, τα νέα δείγματα μπορεί να εμφανίζουν κάποιες τιμές μόνο μία φορά, παραπάνω από μία φορά ή και καθόλου. Στο δεύτερο βήμα υπολογίζεται για κάθε νέο δείγμα η στατιστική που επιθυμείται (π.χ ο μέσος όρος του πληθυσμού) και δημιουργείται μια εκτίμηση της κατανομής της στατιστικής, η Bootstrap κατανομή. Τέλος, χρησιμοποιώντας την Bootstrap κατανομή λαμβάνονται πληροφορίες για το σχήμα, το κέντρο, το τυπικό σφάλμα και την τυπική απόκλιση της δειγματικής κατανομής.

Υπάρχουν διάφορες παραλλαγές της μεθόδου Bootstrap όπως η Μπεϋζιανή Bootstrap, η Ομαλοποιημένη μέθοδος, η Παραμετρική μέθοδος, η Double Bootstrap και η M-out-of-N Bootstrap. Στην Μπεϋζιανή Bootstrap οι παρατηρήσεις που επιλέγονται με δειγματοληψία και επανάθεση δεν έχουν όλες $\frac{1}{n}$ πιθανότητα να

επιλεγούν , παρά κάθε x^i έχει και διαφορετική πιθανότητα. Η μέθοδος είναι αρκετά περιοριστική και δεν προτιμάται για γενική συμπερασματολογία. Στην Ομαλοποιημένη μέθοδο υπολογίζεται μια «ομαλή εκδοχή» της εκτιμήτριας της δειγματικής κατανομής, με τη χρήση μεθόδων εξομάλυνσης πυρήνων. Η Παραμετρική μέθοδος bootstrap χρησιμοποιεί ένα παραμετρικό μοντέλο στο στάδιο της δειγματοληψίας. Είναι σημαντική σαν μέθοδος όταν η παραμετρική κατανομή είναι δύσκολο να εξαχθεί. Στην M-out-of-N μέθοδο επιλέγονται bootstrap δείγματα με μικρότερο μέγεθος από το αρχικό δείγμα(m<n), ώστε να υπάρχει μικρότερη εξάρτηση και η διασπορά της εκτιμήτριας να προσομοιάζεται καλύτερα. Αυτή η μέθοδος είναι βέλτιστη σε περιπτώσεις χρονοσειρών. Η double bootstrap απαιτεί μεγάλη υπολογιστική ισχύ. Δημιουργούνται B^2 bootstrap δείγματα, όπου B είναι τα bootstrap δείγματα που λαμβάνονται από το αρχικό δείγμα αλλά και αυτά που χρησιμοποιούνται για προσαρμογή στην εκτιμήτρια.

Από τη χρήση της bootstrap κατανομής, υπολογίζονται τα διαστήματα εμπιστοσύνης. Ανάλογα με το αν υπάρχει μεροληψία ή/και έλλειψη συμμετρίας στην κατανομή επιλέγεται και η μέθοδος εύρεσης των διαστημάτων εμπιστοσύνης. Οι πιο γνωστές μέθοδοι είναι η percentile, η bias corrected και η μέθοδος με κανονική προσέγγιση. Η **percentile** μέθοδος χρησιμοποιεί τα ποσοστιαία σημεία της κατανομής bootstrap μιας στατιστικής συνάρτησης. Έστω θ η παράμετρος που μας ενδιαφέρει, $\hat{\theta}$ η εκτιμήτρια συνάρτηση και $\hat{\theta}^*(b)$ η τιμή της στατιστικής συνάρτησης για το b bootstrap δείγμα. Συνεπώς σε άπειρες επαναλήψεις το 1-α percentile διάστημα εμπιστοσύνης θα είναι το $[\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}]$ ενώ σε B επαναλήψεις θα είναι το $[\hat{\theta}^*_{(\frac{\alpha}{2}*B)}, \hat{\theta}^*_{((1-\frac{\alpha}{2})*B)}]$ (Efron B. , 1993) . Η μέθοδος **bias corrected** ρυθμίζει τη μεροληψία στη bootstrap κατανομή, αποτελεί καλύτερη εκδοχή των percentile διαστημάτων εμπιστοσύνης και στην πράξη προτιμάται. Για το διάστημα εμπιστοσύνης ισχύει $[\hat{\theta}^*_{(\alpha 1)}, \hat{\theta}^*_{(\alpha 2)}]$ όπου $\alpha_1 = \Phi(\hat{Z}o + \frac{\hat{Z}o + Z_{\alpha/2}}{1-\hat{\alpha}(\hat{Z}o + Z_{\alpha/2})})$ και $\alpha_2 = \Phi(\hat{Z}o + \frac{\hat{Z}o + Z_{1-\alpha/2}}{1-\hat{\alpha}(\hat{Z}o + Z_{1-\alpha/2})})$ με Zα να είναι το 100^{α} ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής και Φ η τυποποιημένη κανονική αθροιστική

κατανομή.(Dicicio T. , 1996) Στη μέθοδο με **κανονική προσέγγιση**, έστω πάλι ότι θ η παράμετρος που μας ενδιαφέρει με κατανομή F και $\hat{\theta}$=T* η εκτιμήτρια συνάρτηση με κατανομή F_T. Για να ισχύει $P(c_{1-a/2} < \hat{\theta} - \theta < c_{a/2}) = 1-a$ τότε το διάστημα εμπιστοσύνης θα είναι $(\hat{\theta} - c_{a/2}, \hat{\theta} - c_{1-a/2})$. Αντικαθιστώντας την F με $\hat{F}$ την T με T* και ύστερα από τους αντίστοιχους υπολογισμούς καταλήγει το διάστημα εμπιστοσύνης να είναι $[\hat{\theta} - (T^*_{(B(1-\alpha/2))} - \theta_{\hat{F}}), \hat{\theta} - (T^*_{(B\alpha/2)} - \theta_{\hat{F}})]$. Πλεονεκτεί σαν μέθοδος, στο ότι παρέχει ένα απλό τρόπο υπολογισμού των διαστημάτων εμπιστοσύνης, ειδικά αν μελετάται η μέση τιμή.

<u>Μέθοδοι διόρθωσης του p-value</u>

Η διόρθωση **Bonferroni** χρησιμοποιείται όταν αρκετές εξαρτημένες ή ανεξάρτητες στατιστικές δοκιμές διεξάγονται ταυτόχρονα. Μετατρέπει στα τεστ ελέγχου το επίπεδο σημαντικότητας σε a=a/n (όπου n=το σύνολο δειγμάτων). Εάν ένα γονίδιο έχει p-value< a/n τότε θεωρείτε στατιστικά σημαντικό. Με αυτή τη μέθοδο όμως λανθασμένα απορρίπτονται πολλά αποτελέσματα(Eye, 2003). Η μέθοδος **FDR** είναι ισχυρότερη από την Bonferroni διότι μπορεί και ξεχωρίζει ποια γονίδια θεωρήθηκαν λανθασμένα μη-στατιστικά σημαντικά. Η μέθοδος είναι η εξής : τα p-value των γονιδίων ταξινομούνται σε αύξουσα σειρά από i=1…n, έστω ότι το επίπεδο σημαντικότητας είναι a, συγκρίνω κάθε p-value ξεκινώντας από το τελευταίο με την τιμή (a*i)/n, αν βρεθεί p με τιμή μικρότερη από αυτή ορίζεται ως η νέα τιμή a. Απορρίπτω τώρα όσα p-value< a(Benjamini & Hochberg). Η μέθοδος **Holm** είναι παρόμοια με τη Bonferroni αλλά πιο ισχυρή. Η μέθοδος είναι η εξής : τα p-value των γονιδίων ταξινομούνται σε αύξουσα σειρά, έπειτα κάθε p(i) συγκρίνεται με την τιμή a=a/(n-i+1) όπου i=1…n και n ο συνολικός αριθμός των γονιδίων του πειράματος. Όσα p-value είναι μικρότερα του επιπέδου σημαντικότητας a, θεωρούνται στατιστικά σημαντικά.(Aickin & Gensler, 1996) Η διόρθωση **Holland** θεωρείται παραλλαγή της μεθόδου Holm. Οι τιμές των p-values των γονιδίων ταξινομούνται από το μικρότερο στο μεγαλύτερο i=1….n. Σημαντικά είναι τα γονίδια εκείνα που έχουν p(i)<a όπου a=1-(1-a)1/(n-i+1)(Holland & Copenhaver, 1988). Η μέθοδος **Sidak** χρησιμοποιείται στους πολλαπλούς ελέγχους. Είναι μια απλή μέθοδος και προτιμάται από τη μέθοδο Bonferroni. Το επίπεδο σημαντικότητας μετατρέπεται σε a=1-(1-a)*1/m και τα γονίδια με p-value<a θεωρούνται στατιστικά σημαντικά (Sidak, 1967).

1.2.5. Ομαδοποίηση

Υπάρχουν αρκετοί αλγόριθμοι ομαδοποίησης που χωρίζουν τα γονίδια σε ομάδες ανάλογα με τα επίπεδα έκφρασής τους. Οι αλγόριθμοι χωρίζονται σε ιεραρχικοί-μη ιεραρχικοί, διαχωριστικοί-συσσωρευτικοί, επιβλέπων-μη επιβλέπων.

Απόσταση: Σημαντικός παράγοντας στην ομαδοποίηση είναι και ο τρόπος μέτρησης της απόστασης, μεταξύ των ομάδων. Οι πιο γνωστοί τύποι αποστάσεων είναι:

Ευκλείδια Απόσταση $d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Απόσταση Manhattan $d_1 = \sum_{i=1}^{n} |x_i - y_i|$ οπού x_i y_i είναι οι τιμές έκφρασης για τα γονίδια X και Y και n είναι τα πειράματα.

Ιεραρχικοί Μέθοδοι Ομαδοποίησης

Είναι μια απλή μέθοδος. Αρχικά κάθε στοιχείο(γονίδιο) αποτελεί μια ξεχωριστή ομάδα, υπολογίζονται οι αποστάσεις μεταξύ των ομάδων και με βάση αυτή την απόσταση τα κοντινότερα στοιχεία συνενώνονται σε νέες ομάδες. Η διαδικασία επαναλαμβάνεται και τερματίζει όταν όλα τα στοιχεία ανήκουν σε μια ομάδα. Οι ιεραρχικοί αλγόριθμοι που χρησιμοποιούνται σε δεδομένα μικροσυστοιχιών είναι οι ακόλουθοι και διαφέρουν στον τρόπο υπολογισμού της απόστασης.

Μέθοδος απλής σύνδεσης. Η απόσταση μεταξύ των ομάδων υπολογίζεται ως η ελάχιστη απόσταση μεταξύ των μελών των ομάδων.

Μέθοδος πλήρους σύνδεσης. Η απόσταση υπολογίζεται ως η μέγιστη απόσταση μεταξύ των μελών των ομάδων.

Μέθοδος σύνδεσης μέσων τιμών. Η απόσταση υπολογίζεται από τις μέσες τιμές.

Μέθοδος σύνδεσης κέντρων. Η απόσταση είναι ίση με την ευκλείδεια απόσταση των κέντρων των ομάδων.

Διαχωριστικοί Μέθοδοι Ομαδοποίησης

Μέθοδος K μέσων (K-means). Πρέπει εξαρχής να οριστεί ο αριθμός των

ομάδων(K) που θα δημιουργηθούν και θα διαμοιραστούν τα δεδομένα. Τα στοιχεία τοποθετούνται στις ομάδες που βρίσκονται πλησιέστερα ως προς αυτά. Υπολογίζεται η απόσταση κάθε στοιχείου από το κέντρο όλων των ομάδων, και ανατίθενται εκ νέου σε πλησιέστερες ομάδες. Γενικά είναι ένας αλγόριθμος ευαίσθητος στις αρχικές συνθήκες οι οποίες παράγουν διαφορετικά αποτελέσματα ομαδοποίησης.

Μέθοδος Αυτό-οργανούμενων χαρτών (SOM). Είναι μια διαχωριστική μέθοδος που βασίζεται στα νευρωνικά δίκτυα και προτάθηκε από τον Kohonen το 1980. Τα νευρωνικά δίκτυα μαθαίνουν να αντιμετωπίζουν το εκάστοτε πρόβλημα από τα δεδομένα τους και τις εξόδους που αντιστοιχούν σε αυτά. Ο αλγόριθμος αυτός περιλαμβάνει το επίπεδο εισόδου και το επίπεδο ανταγωνιστικών νευρώνων με ένα διάνυσμα βαρών ο καθένας. Όταν εφαρμόζεται κάποια είσοδος, οι νευρώνες ανταγωνίζονται και νικητής είναι αυτός του οποίου το διάνυσμα βαρών ομοιάζει περισσότερο με την είσοδο. Πριν να χωριστούν τα γονίδια σε ομάδες τα διανύσματα εκπαιδεύονται(Dalton, Ballarin, & Brun, 2009).

Μέθοδος PCA. Ερευνά δεδομένα πολλών διαστάσεων και επιτρέπει στα σύνθετα δεδομένα να απεικονίζονται σε μειωμένο χώρο διατηρώντας την απόκλισή τους. Από μόνη της είναι δύσκολο να ταξινομήσει τα γονίδια, όμως συνδυαστικά με την K-means ισχυροποιείται.

Μέθοδοι Επιβλεπόμενης Μάθησης

Μέθοδος SVM. Οι διανυσματικές μηχανές υποστήριξης χρησιμοποιούν ένα αρχικό αριθμό δεδομένων για εκπαίδευση, και έπειτα είναι σε θέση να ξεχωρίσουν τα υπόλοιπα δεδομένα με βάση την έκφρασή τους. Προσπαθεί να τα ταξινομήσει σε δυο κατηγορίες, (θετικό σύνολο εκπαίδευσης ή όχι) με τη συνάρτηση kernel.

Μέθοδος Κ κοντινότερων γειτόνων (K-nearest neighbors). Είναι αποτελεσματικός σε μεγάλα σύνολα δεδομένων και έχει αντοχή στα λάθη μέτρησης. Πρέπει εξαρχής να οριστεί ο αριθμός των κοντινότερων γειτόνων (K). Υπολογίζεται η απόσταση του άγνωστου διανύσματος από όλα τα αρχικά διανύσματα, ταξινομούνται οι αποστάσεις και επιλέγονται οι Κ κοντινότερες, βρίσκεται η κλάση που αντιστοιχεί στους Κ κοντινότερους γείτονες και θεωρείται σαν κλάση του άγνωστου διανύσματος.

1.3. ΜΕΤΑ-ΑΝΑΛΥΣΗ

Μετα-ανάλυση είναι η στατιστική προσέγγιση για την σύνθεση αποτελεσμάτων από ανεξάρτητες έρευνες που ερευνούν το ίδιο αντικείμενο. Πρώτος ανέφερε αυτόν τον όρο ο Glass στην ψυχολογία το 1976. Αρχικά χρησιμοποιήθηκε ως μεθοδολογία σε έρευνες της ψυχολογίας, αργότερα των κοινωνικών επιστημών, της οικονομίας και πλέον και στις ιατρικές έρευνες.

1.3.1. Βήματα μετα-ανάλυσης

Θεμελίωση του ερευνητικού ερωτήματος

Πρέπει να είναι διατυπωμένο με σαφήνεια το ερώτημα που ερευνά η μετα-ανάλυση καθώς και οι στόχοι που αναμένεται να επιτευχθούν.

Βιβλιογραφική ανασκόπηση και επιλογή των μελετών

Μέσα από τις ηλεκτρονικές βάσεις δεδομένων (Pubmed, Cochrane library, Scopus, Google Scholar and more) και αρχεία βιβλιοθηκών, επιλέγονται οι μελέτες που περιέχουν τα κατάλληλα δεδομένα. Κάποια από τα βασικά κριτήρια εισαγωγής μελετών είναι το μέγεθος δείγματος, η πληρότητα των πληροφοριών και το έτος δημοσίευσης. Προσοχή πρέπει να δοθεί στην μεροληψία (μελέτες με σημαντικά αποτελέσματα είναι πιο πιθανό να δημοσιευθούν, ολοκληρώνεται η ερευνά τους γρηγορότερα από το προκαθορισμένο, δημοσιεύονται κυρίως στα αγγλικά, χρησιμεύουν στις αναφορές άλλων μελετών). Επίσης γίνεται καταγραφή των μελετών που αποκλείστηκαν και οι αιτίες αποκλεισμού.

Καταγραφή δεδομένων

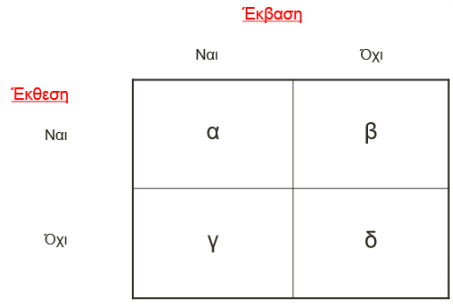
Αφού συγκεντρωθούν όλες οι μελέτες που αφορούν την έρευνα, καταγράφονται οι βασικές τους πληροφορίες σε πίνακες του προγράμματος excel. Περιέχονται πληροφορίες όπως : ο τίτλος του επιστημονικού περιοδικού, τα ονόματα των συγγραφέων, ο αριθμός των συμμετεχόντων(πιο συγκεκριμένα ασθενών-μαρτύρων), η ηλικία και το φύλο τους, το είδος της μελέτης και η έκβαση των αποτελεσμάτων.

Ερμηνεία των αποτελεσμάτων

Ανάλογα με τον τύπο δεδομένων που έχουν οι μελέτες(διχοτομικές

μεταβλητές, συνεχείς μεταβλητές) επιλέγεται και το μέγεθος επίδρασης που θα γίνει η μετα-ανάλυση.

Για **διχοτομικές μεταβλητές** επιλέγεται ο σχετικός κίνδυνος (RiskRatio), ο λόγος αναλογιών (OddsRatio) και η διαφορά κινδύνου (RiskDifference). Η διαφορά κινδύνου εκφράζει τον επιπλέον κίνδυνο να νοσήσει ένα άτομο σε σύγκριση με ένα άλλο που δεν έχει εκτεθεί σε έναν παράγοντα [γ/(γ+δ)]-[α/(α+β)]. Ο σχετικός κίνδυνος εκφράζει την πιθανότητα ενός ατόμου να νοσήσει έχοντας εκτεθεί σε κάποιο παράγοντα σε σχέση με κάποιον που δεν έχει εκτεθεί [α/(α+β)]/[γ/(γ+δ)]. Ο λόγος αναλογιών είναι η πιθανότητα εμφάνισης μιας ασθένειας σε σχέση με την μη εμφάνιση [(α/γ)/(β/δ)] .



Εικόνα 1.10 πείραμα διχοτομικών μεταβλητών. Έκθεση σε παράγοντα. Έκβαση ασθένειας

Για **συνεχείς μεταβλητές** επιλέγεται η τυποποιημένη διαφορά μέσων τιμών και η διαφορά μέσων τιμών μεταξύ των δειγμάτων ελέγχου και αναφοράς. Η διαφορά μέσων τιμών χρησιμοποιείται μόνο όταν όλες οι μελέτες της ανάλυσης χρησιμοποιούν την ίδια κλίμακα αποτελέσματος και είναι ευρέως γνωστή. Για κάθε μελέτη ισχύει $D = \bar{X}_1 - \bar{X}_2$ όπου $\bar{X}_1$ η μέση τιμή των υγιών και $\bar{X}_2$ η μέση τιμή των ασθενών.

Εάν για την τυπική απόκλιση ισχύει $\sigma = \sigma_1 = \sigma_2$ τότε η διακύμανση ορίζεται ως $V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{pooled}$ με $S^2_{pooled} = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$ και το τυπικό σφάλμα είναι $\sqrt{V_D}$.

Εάν για την τυπική απόκλιση ισχύει $\sigma_1 \neq \sigma_2$ τότε η διακύμανση ορίζεται ως $V_D \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ και το τυπικό σφάλμα είναι $\sqrt{V_D}$.

Αντίθετα η τυποποιημένη διαφορά μέσων τιμών μετατρέπει όλα τα μεγέθη

επίδρασης των μελετών σε ένα κοινό μετρικό σύστημα, ώστε να περιλαμβάνονται διαφορετικές μετρήσεις αποτελέσματος στην ίδια ανάλυση. Για κάθε μελέτη ισχύει $d = \frac{\bar{X}_1 - \bar{X}_2}{S_{within}}$ όπου $S_{within} = \sqrt{\frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{n_1 + n_2 - 2}}$. Η διακύμανση ορίζεται ως $V_D = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$ και το τυπικό σφάλμα είναι $\sqrt{V_D}$. Ακόμα και αν ισχύει σ $=\sigma_1=\sigma_2$, υπολογίζεται ο εκτιμητής της τυπικής απόκλισης S_{within} γιατί είναι απίθανο η τυπική απόκλιση των υγιών και ασθενών στη μελέτη (S_1, S_2) να είναι ταυτόσημες (Normand, 1999).

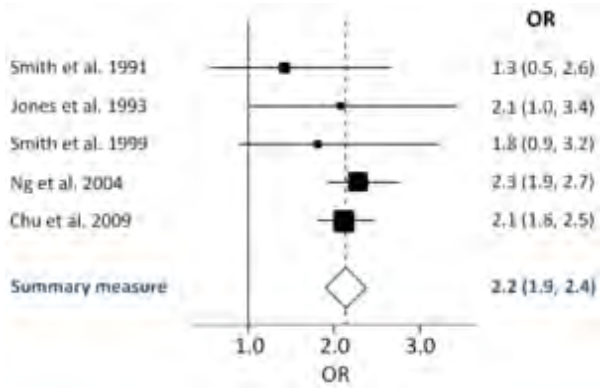
Συνδυαστικά με τα μέτρα επίδρασης των μελετών επιλέγεται και το μοντέλο στατιστικής ανάλυσης που θα εφαρμοστεί. Αν τα μέτρα επίδρασης μεταξύ των μελετών είναι κοινά τότε χρησιμοποιείται το μοντέλο σταθερών επιδράσεων(fixed effects), ενώ αν σε κάθε μελέτη υπάρχει διαφορετικό μέγεθος επίδρασης τότε χρησιμοποιείται το μοντέλο τυχαίων επιδράσεων(random effects model). Στο μοντέλο σταθερών επιδράσεων τα αποτελέσματα της μετα-ανάλυσης αφορούν μόνο των πληθυσμό των μελετών που χρησιμοποιήθηκαν, ενώ στο μοντέλο τυχαίων επιδράσεων οι μελέτες είναι τυχαίο δείγμα και επιτρέπεται η γενίκευση των αποτελεσμάτων (Borenstein, Hedges, Higgins, & Rothstein, 2009). Για τον τύπο της μετα-ανάλυσης σε random effects model ισχύει, $\theta(\tau)_{MLE} = \frac{\Sigma_i W_{i(\tau)} Y_i}{\sum_\kappa W_{i(\tau)}}$ με $W_{i(\tau)} = \frac{1}{s^2{}_i + \tau^2}$ και $Y_i = \bar{x}_{1i} - \bar{x}_{2i}$.

Η ακρίβεια και η εγκυρότητα μιας μετα-ανάλυσης εξαρτώνται σημαντικά από το βαθμό στον οποίο οι επιμέρους μελέτες είναι αρκετά ομοιογενείς μεταξύ τους. Η διαπίστωση ετερογένειας γίνεται με τις γραφικές forest plot, με το Q test και I^2 test και εξετάζει τη μηδενική υπόθεση ότι όλες οι μελέτες αξιολογούν το ίδιο αποτέλεσμα. Το Q τεστ υπολογίζεται αθροίζοντας το τετράγωνο της απόκλισης των εκτιμήσεων κάθε μελέτης από την εκτίμηση μετα-ανάλυσης, σταθμίζοντας τη συνεισφορά κάθε μελέτης όπως στη μετα-ανάλυση. Το I^2 test περιγράφει το ποσοστό της συνολικής διακύμανσης μεταξύ των μελετών που οφείλεται στην ετερογένεια παρά στην τύχη. Υπολογίζεται ως [I^2 =100%*(Q - df)/Q] όπου Q το τεστ Cochrane.

Εάν το I^2 <=25% τότε οι μελέτες θεωρούνται ομογενείς, ενώ αν I^2 >=75% τότε ετερογένεια είναι πολύ μεγάλη. Με τα δενδρογράμματα είναι εφικτή η οπτική ανάλυση ετερογένειας. Εάν οι οριζόντιες γραμμές του γραφήματος, που δείχνουν το διάστημα εμπιστοσύνης, είναι μακριές τότε υπάρχει ετερογένεια(Higgins, Thompson, Deeks, & Altman, 2003). Η ετερογένεια στην παρούσα μελέτη υπολογίστηκε από το στατιστικό πακέτο STATA με την εντολή metan.

Παρουσίαση των αποτελεσμάτων

Ο πιο κοινός τρόπος παρουσίασης των αποτελεσμάτων είναι με δενδρογράμματα (γραφικές forest plot). Παρουσιάζεται το μέγεθος επίδρασης και το διάστημα εμπιστοσύνης κάθε μελέτης ξεχωριστά. Αριστερά σε μια στήλη παρουσιάζονται τα ονόματα των μελετών με χρονολογική σειρά. Το γράφημα αναπαριστά για κάθε μελέτη το μέγεθος επίδρασης(τετράγωνο σχήμα), το διάστημα εμπιστοσύνης(οριζόντια γραμμή) και το γενικό μέτρο επίδρασης της μετα-ανάλυσης(διακεκομμένη κάθετη γραμμή-διαμάντι) (Ried, 2006).



Εικόνα 1.11 Αναπαράσταση ενός δενδρογράμματος για 5 μελέτες και μέθοδο επίδρασης (OR)

1.3.2. Μέθοδοι μετα-ανάλυσης

Η μετα-ανάλυση μικροσυστοιχιών για τον εντοπισμό διαφορικά εκφραζόμενων γονιδίων είναι μία διαδεδομένη τεχνική. Υπάρχουν τέσσερεις μέθοδοι που συνδυάζουν τις γονιδιακές πληροφορίες και τις αναλύουν, η μέθοδος των p-values, η μέθοδος των μεγεθών επίδρασης και η μέθοδος των ranks.

Η μέθοδος συνδυασμού των **p-values** είναι αρκετά εύχρηστη καθώς μπορεί να χρησιμοποιηθεί ανάμεσα σε διαφορετικού είδους μεταβλητές και είναι κατανοητή.

Αρκετές μέθοδοι όπως η μέθοδος του Fischer, του minP και του Stouffer, οδηγούν σε ασφαλή στατιστική συμπερασματολογία. Απαραίτητο όμως στοιχείο για να χρησιμοποιηθούν οι τιμές των p-values σε μία μετα-ανάλυση, είναι να δίδονται οι τιμές των p-values όλων των γονιδίων που μελετώνται και όχι μόνο μια λίστα εξ αυτών.

Μια επίσης διαδεδομένη μέθοδος στη μετα-ανάλυση μικροσυστοιχιών είναι η χρήση **μεγεθών επίδρασης** και συνηθέστερα της διαφοράς μέσων τιμών. Από τους πρώτους που χρησιμοποίησε αυτές τις μεθόδους ήταν ο Choi et al, που προτείνε τη χρήση της μεθόδου permute για τον υπολογισμό των p-values και την εκτίμηση της τιμής FDR. Υπό την προϋπόθεση ότι τα μεγέθη επίδρασης των μελετών μπορούν να συνδυαστούν, επιλέγεται και το κατάλληλο μοντέλο μετα-ανάλυσης. Τα πιο γνωστά μοντέλα είναι το random effect και fixed effect model. Στο fixed effect μοντέλο, όλες οι μελέτες πρέπει να έχουν το ίδιο πραγματικό μέγεθος επίδρασης, ενώ στο random effect να ποικίλει θεωρώντας πως οι μελέτες αποτελούν τυχαίο δείγμα μεγεθών επίδρασης.

Στα πειράματα μικροσυστοιχιών μελετώνται χιλιάδες γονίδια. Ο θόρυβος της πειραματικής διαδικασίας μπορεί να επηρεάσει τα αποτελέσματα εμφανίζοντας ακραίες τιμές. Η μέθοδος των βαθμών κατάταξης **(rank)** τείνει να μην επηρεάζεται από αυτό το φαινόμενο καθώς υπολογίζει το βαθμό διαφορικής έκφρασης κάθε γονιδίου κάθε μελέτης. Αφότου υπολογιστεί ο μέσος όρος ή το άθροισμα ή το γινόμενο των rank όλων των μελετών, χρησιμοποιείται για τη στατιστική ανάλυση. Η στατιστική σημαντικότητα μπορεί να ελεγχθεί περαιτέρω με permutation τεστ. (George Tseng 2012).

Πλεονεκτήματα μετα-ανάλυσης

Η μέθοδος της μετα-ανάλυσης πλεονεκτεί ως προς το ότι δεν μεροληπτεί στην ανάλυση των ξεχωριστών μελετών, δίνει διαφορετική βαρύτητα στην κάθε μελέτη και υπερτερεί της συστηματικής ανασκόπησης. Αυξάνει τη δυνατότητα γενίκευσης των αποτελεσμάτων. Ανιχνεύει και συνυπολογίζει την ετερογένεια. Όσες περισσότερες μελέτες χρησιμοποιεί τόσο στατιστικά σημαντικότερο είναι το αποτέλεσμα. Είναι οικονομική σαν μέθοδος αφού χρησιμοποιεί προϋπάρχουσες μελέτες. Επίσης μπορεί να υπάρξει πρακτική εφαρμογή των αποτελεσμάτων της (Adaikalavan Ramasamy, 2008).

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1. Ερευνητικό ερώτημα

Στην παρούσα εργασία, πραγματοποιήθηκε μετα-ανάλυση μικροσυστοιχιών DNA για την μελέτη της γονιδιακής έκφρασης καρδιαγγειακών παθήσεων. Πιο συγκεκριμένα ερευνάται η νόσος του εμφράγματος του μυοκαρδίου(myocardial disease) και στόχος είναι να εντοπιστούν τα γονίδια που συσχετίζονται με τον κίνδυνο εμφάνισης της ασθένειας.

2.2. Συλλογή δεδομένων

Πραγματοποιήθηκε αναζήτηση βιβλιογραφίας στην επιστημονική βάση δεδομένων Pubmed και στην διεθνή βάση μικροσυστοιχιών GEO. Οι λέξεις-κλειδιά που χρησιμοποιήθηκαν για την αναζήτηση στην Pubmed ήταν "microarray" AND("myocardial ischaemia" OR "myocardial infarction"), ενώ για την GEO η λέξη-κλειδί ήταν "myocardial infarction". Επιλέχθηκαν μελέτες ασθενών μαρτύρων ενώ απορρίφθηκαν όσα άρθρα δεν ήταν σχετικά με το θέμα της εργασίας, όσα επεξεργάζονταν δείγμα microRNA και όσα μελετούσαν δείγματα ζώων. Τα επιλεγμένα άρθρα μελετήθηκαν πλήρως και κατεγράφησαν οι βασικές τους πληροφορίες σε λογιστικά φύλλα excel όπως ο κωδικός της μελέτης, το όνομα της εργασίας, το δείγμα ιστού που χρησιμοποιήθηκε, η μέθοδος ανάλυσης και ο αριθμός των προς μελέτη γονιδίων.

2.3 Καταγραφή δεδομένων

Ενώ όλες οι μελέτες από Pubmed παρείχαν πληροφορίες για τον αριθμό των γονιδίων που εξέτασαν και για το πλήθος των ασθενών-υγιών καμία δεν έδινε τα πραγματικά δεδομένα, δηλαδή το συγκεντρωτικό αρχείο με τις τιμές έκφρασης του κάθε γονιδίου για κάθε εξεταζόμενο ξεχωριστά. Ένα τέτοιου είδους πρόβλημα οδήγησε στο να αποσυρθούν όλες αυτές οι μελέτες απ ’την μετα-ανάλυση. Αντίθετα όλες οι μελέτες από την GEO παρείχαν τα πλήρη δεδομένα ώστε να συμπεριληφθούν στην έρευνα, εκτός από μία που δεν προσδιόριζε ποιοι είναι ασθενείς και ποιοι μάρτυρες και αφαιρέθηκε.

Οι πληροφορίες που χρησιμοποιήθηκαν από κάθε μελέτη της βάσης δεδομένων GEO είναι ο πίνακας γονιδιακής έκφρασης(series matrix file) και οι πληροφορίες από το αρχείο της πλατφόρμας των μικροσυστοιχιών (GPL). Ο πίνακας δεδομένων γονιδιακής έκφρασης περιλαμβάνει την έκφραση των γονιδίων για κάθε

δείγμα έπειτα από τη χρήση διαφόρων αλγορίθμων κανονικοποίησης ή/και λογαρίθμησης. Επίσης οι πληροφορίες σχετικά με την αντιστοίχιση των ανιχνευτών με τα αντίστοιχα ονόματα των γονιδίων αντλήθηκαν από το αρχείο GPL.

Για την ανάλυση των δεδομένων χρησιμοποιήθηκε ο πίνακας γονιδιακής έκφρασης, όπου έπειτα από επεξεργασία οι γραμμές αναπαριστούσαν τα ονόματα των γονιδίων (GENE SYMBOL) και οι στήλες τις τιμές έκφρασης των γονιδίων για τους υγιείς-ασθενείς (CONTROL-CASE). Αν στην εγγραφή κάποιου ανιχνευτή δεν αντιστοιχιζόταν κάποιο γονίδιο(είτε γιατί δεν έγινε υβριδοποίηση είτε γιατί αφορούσε dark-bright corner), αυτή η εγγραφή διαγράφονταν. Επίσης επειδή πολλοί ανιχνευτές υβριδοποιούνται από πολλαπλά μετάγραφα του ίδιου γονιδίου, εμφανίζονταν εγγραφές για το ίδιο γονίδιο με διαφορετικές τιμές έκφρασης. Για αυτά τα γονίδια δημιουργούταν μια νέα γραμμή με τον μέσο όρο εμφάνισής τους. Στη συνέχεια προκειμένου να διεξαχθεί σωστά η μετα-ανάλυση επιλέχθηκαν μόνο τα γονίδια που υπήρχαν τουλάχιστον σε δύο μελέτες.

C1 | Gene Symbol

	A	B	C	D	E	F	G	H
	ID_REF	ID	Gene Symbol	Case_1	Case_2	Case_3	Case_4	Case_5
	229819_at	229819_at	A1BG	6.461131136	6.461083384	6.246361891	6.566609	6.387993
	232462_s_at	232462_s_a	A1BG-AS	6.370162018	6.188048657	6.248392418	6.221658	6.340023
	220951_s_at	220951_s_a	A1CF	5.82982005	6.027904453	5.826461498	6.050966	5.879504
	232422_at	232422_at	A2LD1	5.594188259	5.070654418	5.475818332	5.397859	5.12119
	237869_at	237869_at	A2LD1	4.770897338	4.761997953	4.316379619	4.586172	4.928754
	1558450_at	1558450_at	A2M	4.877496793	5.330395185	4.703954025	5.258187	4.923066
	217757_at	217757_at	A2M	4.423646838	4.636086339	4.431002141	4.371609	4.425529
	1553505_at	1553505_at	A2ML1	3.726544512	3.950735003	3.580230672	3.742045	3.826824
	1564307_a_at	1564307_a_	A2ML1	3.479119441	3.496684082	3.539887333	3.471531	3.67959
	219488_at	219488_at	A4GALT	7.403459304	7.40258025	7.144505228	7.388372	7.363064
	221131_at	221131_at	A4GNT	5.461358833	5.262581778	5.027248816	5.455039	5.35169
	1562228_s_at	1561490_at	AAA1	6.09950942	6.39937817	5.441623025	6.272811	6.485807
	234117_at	234117_at	AAA1	4.049712852	3.916069144	3.544198373	3.954126	4.224281
	218075_at	218075_at	AAAS	7.451237465	7.646271459	7.068039404	7.470139	7.628543
	218434_s_at	218434_s_a	AACS	6.378817003	6.358230594	6.460965651	6.409497	6.314743
	1570020_at	1570020_at	AACSP1	2.927994644	2.958158644	2.991824384	3.162531	2.995431

Εικόνα 2.1 Παρουσίαση ενός τμήματος του πίνακα γονιδιακής έκφρασης, από τη μελέτη με κωδικό GSE48060. Εμφανίζονται πολλαπλά probes του ίδιου γονιδίου με διαφορετικές τιμές έκφρασης.

A	B	C	D	E	F	G	H
Gene Symbol	Case_1	Case_2	Case_3	Case_4	Case_5	Case_6	Case_
A1BG	6.461131	6.461083	6.246362	6.566609	6.387993	6.56178	6.31(
A1BG_AS	6.370162	6.188049	6.248393	6.221659	6.340024	6.331412	6.16
A1CF	5.82982	6.027905	5.826461	6.050966	5.879504	5.789421	5.94
A2LD1	5.182543	4.916326	4.896099	4.992015	5.024972	4.58729	4.88
A2M	4.650572	4.983241	4.567478	4.814898	4.674298	4.737305	4.59
A2ML1	3.602832	3.72371	3.560059	3.606788	3.753207	3.410562	3.56
A4GALT	7.403459	7.40258	7.144505	7.388372	7.363064	7.24962	7.35
A4GNT	5.461359	5.262582	5.027249	5.455039	5.35169	5.303545	5.23
AAA1	5.074611	5.157723	4.492911	5.113469	5.355044	4.83584	5.11
AAAS	7.451238	7.646271	7.068039	7.47014	7.628543	7.184978	7.12
AACS	6.378817	6.358231	6.460966	6.409497	6.314743	6.427484	6.73
AACSP1	2.927995	2.958159	2.991824	3.162531	2.995431	2.819548	3.03
AADAC	4.270115	4.368345	4.481071	4.394915	4.507827	4.523181	4.32
AADACL2	2.835705	2.851672	2.81446	3.040591	3.025333	2.785988	2.80
AADAT	5.574085	5.704424	5.170699	5.618314	5.678475	5.56283	5.66
AAGAB	6.589942	6.335407	6.789129	6.419102	6.136625	6.858229	6.60

Εικόνα 2.2 Παρουσίαση ενός τμήματος του πίνακα γονιδιακής έκφρασης, από τη μελέτη με κωδικό GSE48060. Η τιμή έκφρασης κάθε γονιδίου υπολογίστηκε από τον μέσο όρο των πολλαπλών ιχνηθετών.

2.4. Στατιστική ανάλυση

Στη μετα-ανάλυση χρησιμοποιούνται διαφορετικές μελέτες που ερευνούν το ίδιο αντικείμενο, συνδυάζονται τα αποτελέσματά τους και εξάγονται συμπεράσματα με μεγαλύτερη στατιστική ακρίβεια. Ανάλογα με το είδος των μεταβλητών που χρησιμοποιούνται και με το αν υπάρχει ετερογένεια μεταξύ των μελετών, επιλέγεται το μέγεθος επίδρασης και το μοντέλο στατιστικής ανάλυσης αντίστοιχα για τη μετα-ανάλυση.

Στην παρούσα εργασία επιλέχθηκε το μοντέλο τυχαίων επιδράσεων καθώς οι επιμέρους μελέτες δίνουν διαφορετική βαρύτητα στο μέτρο επίδρασης και δεν έχουν την ίδια επίδραση στον πληθυσμό. Επιπλέον εφόσον οι μεταβλητές από κάθε μελέτη ήταν συνεχείς, χρησιμοποιήθηκε ως μέγεθος επίδρασης η τυποποιημένη διαφορά μέσων τιμών(για τους ασθενείς και τους υγιείς) μέσω του ελέγχου t. Για τον τύπο του t-test ισχύει $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$ (όπου $\bar{x}_1 - \bar{x}_2$ = διαφορά μέσων τιμών δείγματος και $\mu_1 - \mu_2$ = διαφορά μέσων τιμών πληθυσμού). $S_{\bar{x}_1 - \bar{x}_2}$ είναι το τυπικό σφάλμα (standard error) και ισούται με $\sqrt{\frac{S^2_{pooled}}{n_1} + \frac{S^2_{pooled}}{n_2}}$ όπου $S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$. Επειδή όμως στις μελέτες μικροσυστοιχίας είναι σύνηθες να υπάρχει μικρό μέγεθος δείγματος και μη κανονική κατανομή των τιμών

έκφρασης, μετά τον έλεγχο t πραγματοποιήθηκε η μέθοδος bootstrap για την ταυτοποίηση των διαφορικά εκφρασμένων γονιδίων.

Αρχικά κάθε μελέτη εισήχθη σαν πίνακας στο στατιστικό πρόγραμμα STATASE (έκδοση 13), όπου οι στήλες αντιπροσώπευαν τα γονίδια και οι γραμμές τα άτομα της μελέτης. Στο τέλος του πίνακα προστέθηκαν δύο επιπλέον στήλες όπου διευκρινίζουν ποια άτομα είναι ασθενείς-ποια υγιείς (case=1,control=0) και τον αριθμό της μελέτης. Για κάθε μελέτη πραγματοποιήθηκε ο έλεγχος t για κάθε γονίδιο προκειμένου να βρεθούν τα γονίδια που εκφράζονται διαφορικά με εντολή

```
ttest `variable', by(case_control)
```

Στη συνέχεια πραγματοποιήθηκε επιπλέον η κανονική προσέγγιση της μεθόδου bootstrap με την εντολή

```
bootstrap t=r(t), reps(1000) strata(case_control): ttest `var',by(case_control) uneq)
```

και υπολογίστηκαν τα νέα διαστήματα εμπιστοσύνης, οι τιμές του τυπικού σφάλματος και οι τιμές του p-value για κάθε γονίδιο. Εφαρμόστηκαν επιπλέον οι μέθοδοι διόρθωσης του p-value και εντοπίστηκαν τα στατιστικά σημαντικά γονίδια της κάθε μελέτης ξεχωριστά χρησιμοποιώντας τις εντολές

```
multproc, pval(p)  meth(simes)  rej(simes)

multproc, puncor(0.01) pval(p) meth(simes) rej(fdr)

multproc, pval(p) meth(bonferroni) rej(bonf)

multproc, pval(p) meth(sidak) rej(sidak)

multproc, pval(p) meth(holm) rej(holm)

multproc, pval(p) meth(holland) rej(holland)
```

Προκειμένου να πραγματοποιηθεί η μετα-ανάλυση τα δεδομένα των επιμέρους μελετών συνενώθηκαν σε ένα αρχείο. Από τη μετα-ανάλυση, για κάθε γονίδιο υπολογίστηκαν οι τιμές του τυπικού σφάλματος (standard error), του p-value και του z test. Με κριτήριο το p-value θεωρήθηκαν στατιστικά σημαντικά μόνο τα γονίδια που είχαν τιμή $p<0,05$. Γι' αυτά τα γονίδια εφαρμόστηκαν επίσης οι μέθοδοι διόρθωσης του p-value FDR (simes), Bonferroni, Sidak Holm, Holland και παρουσιάστηκαν σε γραφήματα. Επίσης διερευνήθηκαν τα στατιστικώς σημαντικά

γονίδια κατά FDR αλλά και η πιθανή εμφάνιση κοινών γονιδίων που κρίθηκαν σημαντικά σε όλες τις μελέτες. .

2.5. Ερμηνεία αποτελεσμάτων

Με τη χρήση της πλατφόρμας bioCompendium βρέθηκαν οι βιολογικές πληροφορίες και τα βιοχημικά μονοπάτια των στατιστικά σημαντικών γονιδίων, καθώς και οι σχέσεις τους με ασθένειες. Το bioCompendium είναι μια υψηλής απόδοσης πλατφόρμα ανάλυσης δεδομένων, που δέχεται ως είσοδο λίστα γονιδίων ή πρωτεϊνών. Συγκρίνει αποτελέσματα από διαφορετικές πειραματικές συνθήκες, παρουσιάζει αλληλεπιδράσεις μεταξύ των πρωτεϊνών και συντάσσει ομάδες με βάση την ομολογία ακολουθίας των γονιδίων (bioCompendium).

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

Για την κατασκευή γονιδιακού προφίλ για την πρόβλεψη του εμφράγματος του μυοκαρδίου, συλλέχθηκαν μελέτες από επιστημονικές βάσεις δεδομένων. Πιο συγκεκριμένα βρέθηκαν μέσω αναζήτησης 162 άρθρα από τη βάση δεδομένων Pubmed και 4903 μελέτες από τη βάση μικροσυστοιχιών GEO. Μετά από ανασκόπηση των μελετών και εφαρμογή των περιορισμών που αναλύθηκαν στην ενότητα της μεθοδολογίας, ο αριθμός περιορίστηκε σε 7 μελέτες από τη βάση Pubmed και 5 μελέτες από τη βάση GEO. Για τη μετα-ανάλυση περιλήφθησαν μόνο τα άρθρα που περιείχαν τα πλήρη δεδομένα, τα οποία και παρουσιάζονται στον ακόλουθο πίνακα.

Στον ακόλουθο πίνακα παρουσιάζεται ο αριθμός ανιχνευτών και τα συνολικά γονίδια της κάθε μελέτης. Εάν κάποιο γονίδιο εμφανιζόταν μόνο σε μια μελέτη αφαιρούνταν. Τελικά τα συνολικά γονίδια προς μετα-ανάλυση περιορίστηκαν σε 31.180.

GEO ID	PUBMED ID	AUTHORS	CASES	CONTROLS	Πλατφόρμα μικροσυστοιχίας	Συνολικοί ανιχνευτές	Συνολικά γονίδια	Γονίδια προς μετα-ανάλυση
GSE 48060	PMID: 24801707	Suresh R et al.	30	22	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	42450	21049	21037
GSE 60993	PMID: 26025919	Park HJ et al.	7	7	Illumina HumanWG-6 v3.0 expression beadchip	35966	25438	25162
GSE 61144	PMID: 26025919	Park HJ et al.	7	10	Sentrix Human-6 v2 Expression BeadChip	30535	24995	24766
GSE 66360			49	50	Affymetrix Human Genome U133 Plus 2.0 Array	42450	21049	21037

Πίνακας 1. Πληροφορίες μελετών προς μετα-ανάλυση. Ο συνολικός αριθμός ανιχνευτών, ο συνολικός αριθμός γονιδίων και τα τελικά γονίδια της μετα-ανάλυσης.

Προκειμένου να γίνει η μετα-ανάλυση, χρησιμοποιήθηκε το μοντέλο τυχαίων επιδράσεων (random effects model) εφόσον υπήρχε ετερογένεια και επιλέχθηκε ως μέθοδος επίδρασης η τυποποιημένη διαφορά μέσων τιμών, καθώς είναι έγκυρη και πιο συχνά χρησιμοποιούμενη. Για να εκτιμηθεί η κατανομή της μέσης τιμής του δείγματος χρησιμοποιήθηκε η t-στατιστική. Επιπλέον εφόσον υπήρχε μικρό δείγμα και μη κανονική κατανομή, έγινε διόρθωση μέσω της εμπειρικής κατανομής της μεθόδου bootstrap, ώστε να βρεθεί η εκτίμηση της κατανομής της t-στατιστικής χωρίς τη χρήση της κανονικής θεωρίας(t). Με αυτό τον τρόπο υπολογίστηκαν το τυπικό σφάλμα και διαστήματα εμπιστοσύνης μεγαλύτερης ακρίβειας, και η μετα-ανάλυση που ακολούθησε έγινε με κανονική κατανομή των effect size.

Από τη μετα-ανάλυση δημιουργήθηκε ένα αρχείο που δίνει για κάθε γονίδιο, τις τιμές του τυπικού σφάλματος (standard error), του p value και του z test. Τα γονίδια με p-value<0.05 θεωρήθηκαν στατιστικώς σημαντικά, όμως ο μεγάλος αυτός αριθμός γονιδίων φανερώνει την ύπαρξη false positive αποτελεσμάτων. Σε αυτά τα p-values εφαρμόστηκαν οι μέθοδοι διόρθωσης FDR (Simes) – Bonferroni -Holm - Sidak - Holland. Αυτοί οι μέθοδοι διόρθωσης περιόρισαν ακόμα περισσότερο τα στατιστικώς σημαντικά γονίδια, που συγκεντρώνονται στον ακόλουθο πίνακα.

Πίνακας 2. Εμφανίζεται ο αριθμός των γονιδίων, μετά τη μετα-ανάλυση των μελετών, σε p-value<0,05 και σε διορθωμένο p-value ανάλογα με τη μέθοδο διόρθωσης.

	pvalue<0,05	Simes 0,05	FDR 0,01	Bonferroni	Holm	Sidak	Holland
Αριθμός στατιστικά σημαντικών γονιδίων	5.566	1.570	666	163	163	163	164

Επιπλέον για να ενισχυθεί η εγκυρότητα των αποτελεσμάτων της μετα-ανάλυσης, έγινε εφαρμογή της μεθόδου bootstrap και των μεθόδων διόρθωσης του p-value σε κάθε μία μελέτη ξεχωριστά και συγκρίθηκαν τα αποτελέσματά τους με αυτά της μετα-ανάλυσης. Από τη χρήση της μεθόδου bootstrap βρέθηκε ο αριθμός των στατιστικά σημαντικών γονιδίων της κάθε μελέτης και επιλέχθηκαν γονίδια με p-value μικρότερο του 0.05, γονίδια με p-value μικρότερο του 0.01 και γονίδια που σε

επίπεδο σημαντικότητας 1% κρίθηκαν σημαντικά από τη μέθοδο διόρθωσης FDR. Αυτός ο αριθμός γονιδίων συγκρίθηκε με τον αντίστοιχο αριθμό γονιδίων της μετα-ανάλυσης και παρουσιάζεται στον πίνακα 3. Από τον πίνακα αυτόν εμφανίζεται η μέθοδος της μετα-ανάλυσης να εντοπίζει περισσότερα σημαντικά αποτελέσματα από ότι η κάθε μία μελέτη μόνη της.

Στον πίνακα 4 συγκρίθηκαν τα στατιστικώς σημαντικά γονίδια της μετα ανάλυσης, με το άθροισμα των σημαντικών γονιδίων των επιμέρους μελετών (αποκλείοντας τις διπλότυπες τιμές). Μελετήθηκαν σε επίπεδο σημαντικότητας 5% και σε επίπεδο 1% διορθωμένου κατά FDR. Τα αποτελέσματα αυτής της σύγκρισης δείχνουν τις μελέτες αθροιστικά να εντοπίζουν ένα πολύ μεγάλο αριθμό γονιδίων σε σχέση με τη μετα-ανάλυση. Σε ένα τόσο μεγάλο όμως δείγμα ανάλυσης, είναι αναμενόμενο να υπάρχουν πολλά false positive αποτελέσματα, συνεπώς η μετα-ανάλυση είναι πιο αξιόπιστη.

Επιπλέον, υπολογίστηκε και παρουσιάστηκε στον πίνακα 5 για κάθε μελέτη ο αριθμός των ταυτόσημων γονιδίων με τα γονίδια από τη μετα-ανάλυση, και ο αριθμός των κοινών γονιδίων μεταξύ των τεσσάρων μελετών. Σε επίπεδο σημαντικότητας 1% διορθωμένου κατά FDR, η μελέτη GSE48060 έχει (0/3) κοινά σημαντικά γονίδια με τη μετα-ανάλυση, η GSE60993 έχει (22/112) κοινά, η GSE61144 έχει (230/2198) κοινά και η GSE66360 έχει (135/1170) κοινά γονίδια. Αυτό δείχνει ότι η μετα-ανάλυση μπορεί και περιορίζει τα false positive αποτελέσματα. Επιπλέον βρέθηκαν 313 γονίδια που να είναι στατιστικά σημαντικά και στη μετα-ανάλυση αλλά και στις τέσσερεις μελέτες κατά FDR (α=0,01), και βρέθηκαν 2.119 κοινά στατιστικά σημαντικά γονίδια μεταξύ των τεσσάρων μελετών σε επίπεδο σημαντικότητας 5%.

Πίνακας 3. Παρουσιάζεται για κάθε μελέτη αλλά και για τον τελικό πίνακα μετα-ανάλυσης ο συνολικός αριθμός στατιστικά σημαντικών γονιδίων σε επίπεδο σημαντικότητας 0.05 - 0.01, κατά FDR και κατά Holland.

	GSE 48060	GSE 60993	GSE 61144	GSE 66360	Πίνακας γονιδίων που έγινε η μετα-ανάλυση
Αριθμός σημαντικών γονιδίων σε $p<0.05$	2.843	4.250	5.916	4.694	5.566
Αριθμός σημαντικών γονιδίων σε $p<0.01$	830	1.960	3.923	2.698	2.777
Αριθμός σημαντικών γονιδίων κατά FDR	3	112	2.198	1.170	666
Αριθμός σημαντικών γονιδίων κατά Holland.	3	15	268	325	164

Πίνακας 4. Συγκρίνεται ο αριθμός των στατιστικά σημαντικών γονιδίων της μετα-ανάλυσης, με το άθροισμα των στατιστικά σημαντικών γονιδίων των τεσσάρων μελετών (GSE 48060, GSE 60993, GSE 61144, GSE 66360) εφόσον περιορίστηκαν οι διπλότυπες τιμές.

	Αθροιστικός πίνακας γονιδίων των τεσσάρων μελετών	Πίνακας γονιδίων που έγινε η μετα-ανάλυση
Αριθμός σημαντικών γονιδίων σε $p<0.05$	12.149	5.566
Αριθμός σημαντικών γονιδίων κατά FDR	3.166	666

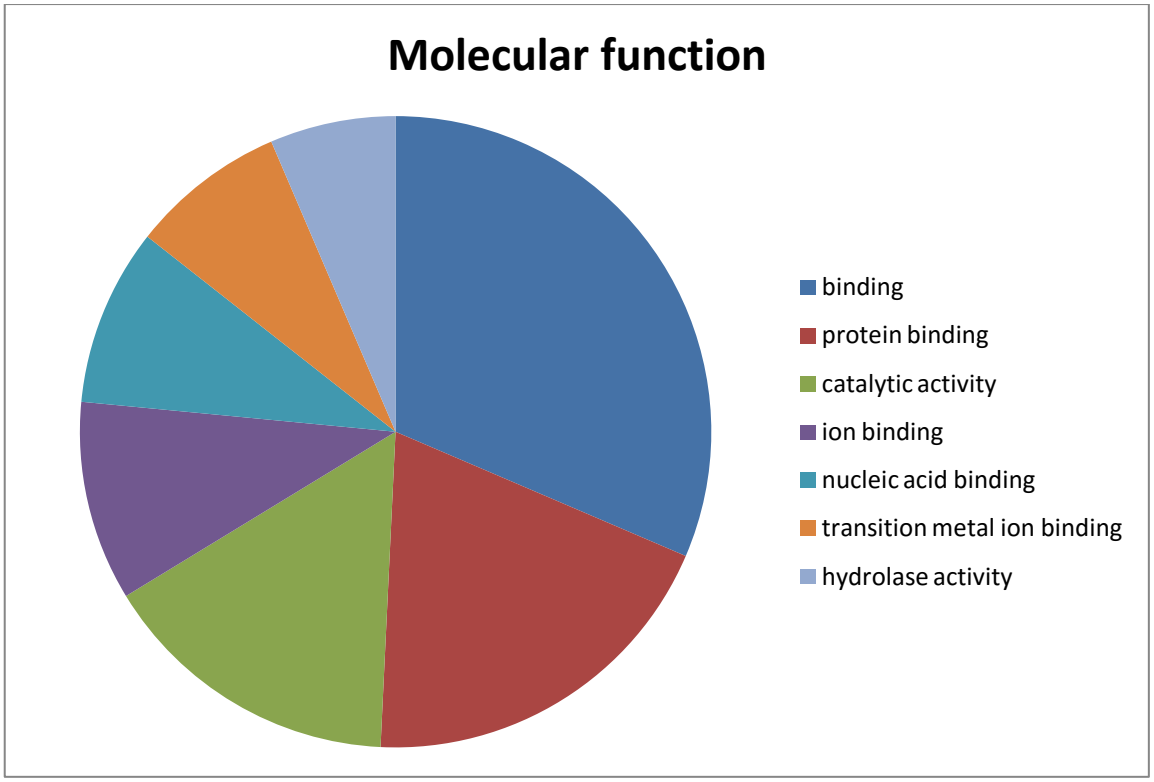
Πίνακας 5. Εμφανίζεται για κάθε μελέτη ο αριθμός των γονιδίων που χαρακτηρίστηκαν στατιστικώς σημαντικά και στη μετα-ανάλυση και στη μελέτη, κατά τη μέθοδο διόρθωσης FDR.

	GSE 48060	GSE 60993	GSE 61144	GSE 66360
Κοινά σημαντικά γονίδια κατά FDR	0	22	230	135

bioCompendium

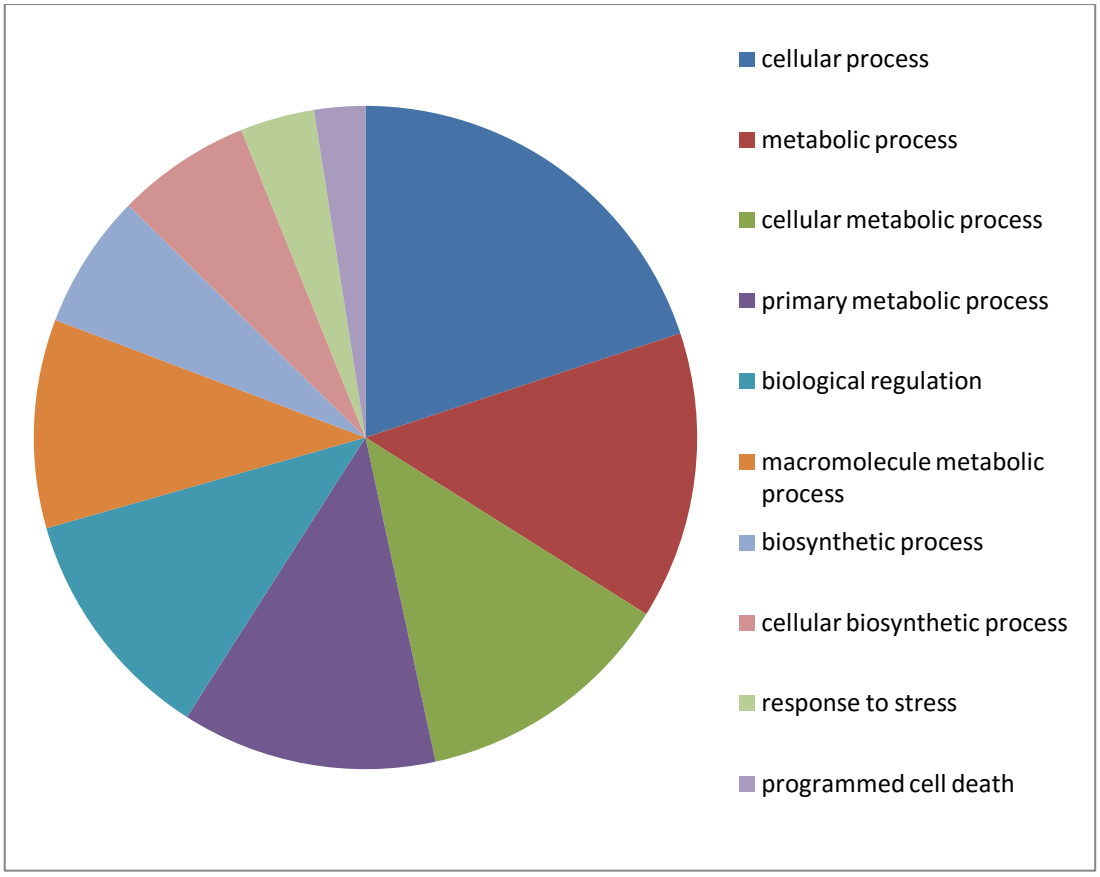
Μοριακή λειτουργία

Θέτοντας ως είσοδο στο bioCompendium τα 470 γονίδια με το μικρότερο p-value που βρέθηκαν να έχουν στατιστικά σημαντική γενετική συσχέτιση με το έμφραγμα του μυοκαρδίου, σύμφωνα με τη μέθοδο διόρθωσης FDR, επεστράφησαν πληροφορίες για τη μοριακή λειτουργία 341 γονιδίων. Πιο συγκεκριμένα 286 από τα 341 γονίδια συμμετέχουν στη δέσμευση, ενώ 176 γονίδια συμμετέχουν στην πρωτεϊνική δέσμευση, 140 γονίδια στην καταλυτική δράση, 94 γονίδια στη δέσμευση ιόντων, 83 γονίδια στη δέσμευση νουκλεϊκού οξέος και στις ακόλουθες λειτουργίες που εμφανίζονται στο γράφημα.



Βιολογικές διεργασίες

Από τα 426 γονίδια, τα 325 συμμετέχουν σε βιολογικές διεργασίες όπως στην κυτταρική διεργασία (cellular process), στην διαδικασία του μεταβολισμού(metabolic process), στον προγραμματισμένο θάνατο κυττάρων (programmed cell death), στην αντίδραση στο στρες (response to stress) και στη μορφογένεση της καρδιάς (heart morphogenesis). Οι κυριότερες λειτουργίες εμφανίζονται στο γράφημα που ακολουθεί.

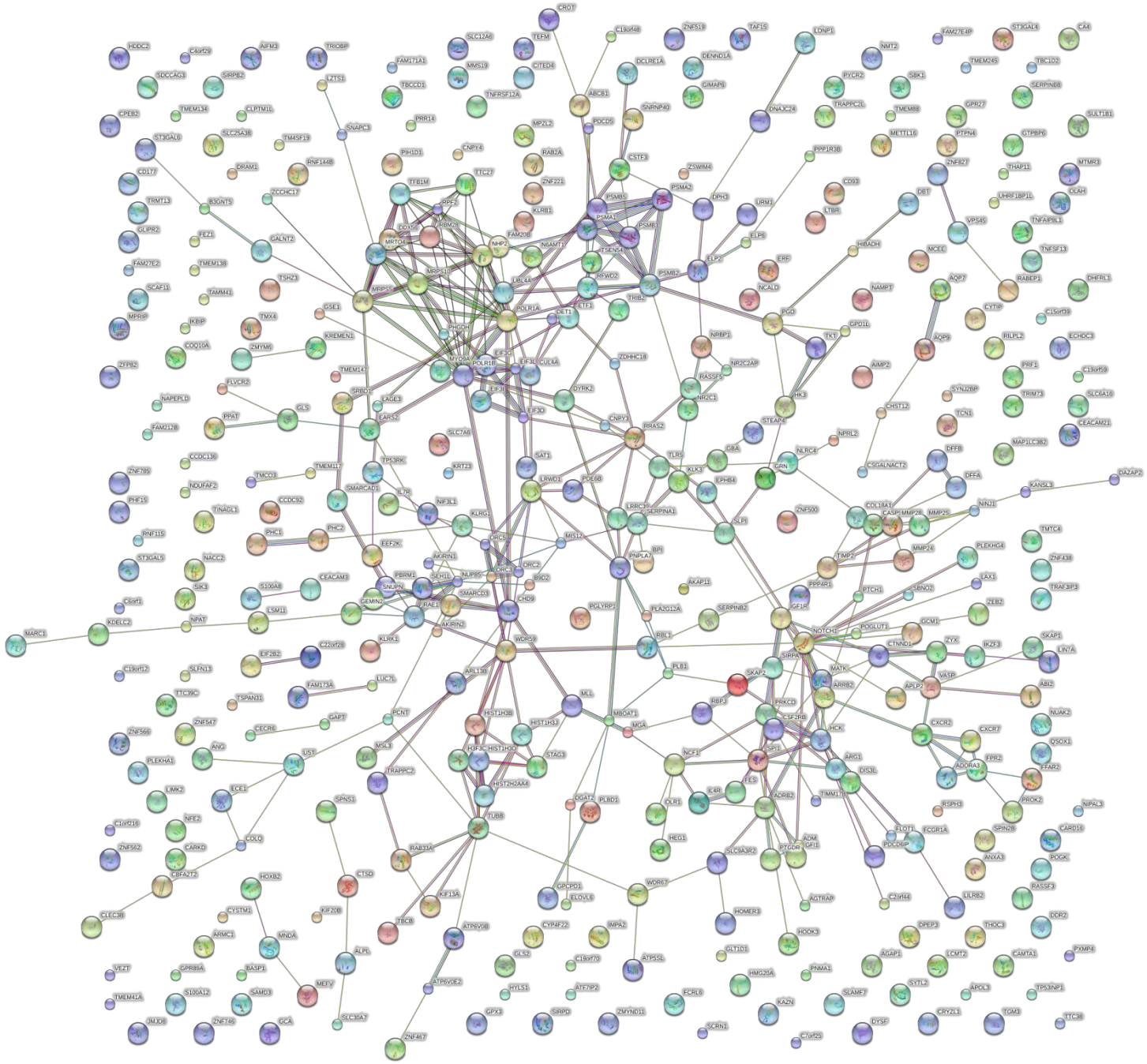


Βιοχημικά μονοπάτια

Επιπλέον με βάση το bioCompendium δόθηκαν πληροφορίες και για τα βιοχημικά μονοπάτια αυτών των γονιδίων. Συνολικά συμμετέχουν σε 118 βιοχημικά μονοπάτια. Από τα γονίδια που παρουσιάζουν στατιστικά σημαντική γενετική συσχέτιση με το έμφραγμα του μυοκαρδίου, υπάρχουν αρκετά που ελέγχουν την αλληλεπίδραση των κυτταροκινών με τους υποδοχείς τους (Cytokine-cytokine receptor interaction), τον τύπου 1 σακχαρώδη διαβήτη (Type I diabetes mellitus), τη νόσο του Πάρκινσον (Parkinson's disease), τον καρκίνο στο πάγκρεας (Pancreatic cancer) και την ιογενή μυοκαρδίτιδα (Viral myocarditis).

Αλληλεπιδράσεις μεταξύ πρωτεϊνών

Στην εικόνα που ακολουθεί παρουσιάζεται το πρωτεϊνικό δίκτυο των (426) γονιδίων και παρατηρούμε 3 κύριες ομάδες γονιδίων που εμφανίζουν μεγάλο αριθμό αλληλεπιδράσεων. Με πιο έντονη γραμμή παρουσιάζονται οι ισχυρότερες συσχετίσεις μεταξύ των γονιδίων. Στο παράρτημα εμφανίζεται η λίστα με τα ονόματά τους.



Στην παρούσα πτυχιακή εργασία πραγματοποιήθηκε μετα-ανάλυση μικροσυστοιχιών DNA στο STATA για την εύρεση γονιδίων που εκφράζονται διαφορικά στη νόσο του εμφράγματος του μυοκαρδίου. Η μετα-ανάλυση ανίχνευσε γονίδια που έχουν στατιστικά σημαντική συσχέτιση με τη νόσο, τα οποία δεν βρέθηκαν να έχουν συσχέτιση στις επιμέρους μελέτες. Ωστόσο χρειάζεται περαιτέρω ανάλυση των γονιδίων αυτών για να αξιολογηθεί η σχέση τους με την νόσο του μυοκαρδίου και των καρδιαγγειακών παθήσεων γενικότερα.

4. ΒΙΒΛΙΟΓΡΑΦΙΑ

Ramasamy Adaikalavan, Mondry A.,et. al. (2008). Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. *PLoS Med*.

Aickin M. & Gensler H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health, v.86*.

Benjamini Yoan & Hochberg Yosef. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*.

bioCompendium. from http://biocompendium.embl.de/

Borenstein Michael, Hedges L. V., et al. (2009). *Introduction to Meta-Analysis*.

Dalton, L., Ballarin, V., & Brun, M. (2009). Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics. *Current Genomics*.

Diciccio T. , E. B. (1996). Bootstrap Cnfidence Intervals. *Statistical Science, 11*.

Efron B. , T. R. (1993). *An introduction to the Bootstrap*.

Eye, A. V. (2003). *Configural Frequency Analysis: Methods, Models, and Applications*: Psychology Press.

Fay, D. S., & Gerow, K. (2013). A biologist's guide to statistical thinking and analysis. *WormBook*.

George Tseng , et al. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ : British Medical Journal, 327(7414)*.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-Type Multiple Testing Procedure. *Psychological Bulletin, 104*.

Løgstrup, S., & O'Kelly, S. (2012). European Cardiovascular Disease Statistics: European Heart Network and European Society of Cardiology.

Mendis, S. (2014). GLOBAL STATUS REPORT on noncommunicable diseases 2014. www.who.int: WORLD HEALTH ORGANIZATION.

Normand, T. (1999). tutorial in biostatistics meta-analysis: formulating,evaluating,combining, and reporting. *STATISTICS IN MEDICINE*.

Ried, K. (2006). Interpreting and understanding meta-analysis graphs. A practical guide. *Australian family physician, 35*(8).

Sidak, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions *Journal of the American Statistical Association, 62*.

Yang, Yee Hwa et al. (2002). Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation. *Nucleic Acids Research*, 30(4), e15.

Trevino V, Falciani F, Barrera-Saldaña HA. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. Mol. Med. 2007;13:527–541. doi: 10.2119/2006-00107.Trevino.

ΣΙΜΟΥ, Ε. (2008). *Εθνικό Σχέδιο Δράσης για τα Καρδιαγγειακά Νοσήματα 2008 - 2012*.

Bumgarner, R. (2013). Overview of DNA Microarrays: Types, Applications, and Their Future. Current Protocols in Molecular Biology. 101:22.1:22.1.1–22.1.11.

4.1. ΠΑΡΑΡΤΗΜΑ

Κώδικας Stata

αντιμετάθεση πίνακα

```
xpose, clear
```

ένωση πινάκων

```
append using "filename.dta"
```

η μία μελέτη να είναι ανοιχτεί στον editor του stata και η άλλη να ορίζεται στο filename.

Κώδικας t-test

```
set more off
file open meta using results.txt, write append
file write meta "gene"
file write meta " , "
file write meta "t"
file write meta " , "
file write meta "r(se)"
file write meta " , "
file write meta "p"
file write meta " , "
file write meta "e(se)" _n


foreach var of varlist a1bg-a1cf {
preserve
qui ttest `var', by(case_control) uneq
file write meta "`var'"
file write meta " , "
file write meta "`r(t)'"
file write meta " , "
file write meta "`r(se)'"
file write meta " , "
file write meta "`r(p)'"

bootstrap t=r(t), reps(1000) strata(case_control): ttest
`var',by(case_control) uneq
mat se=e(se)
```

```
local se=se[1,1]
file write meta " , "
file write meta "`se'" _n

restore

}

file close meta
```

Κώδικας μετα-ανάλυσης bootstrap

```
set more off
sum gene_num
local x=r(max)
file open meta using metanresults2.txt,write append

file write meta "gene"
            file write meta " "
            file write meta "diff"
            file write meta " "
            file write meta "se"
            file write meta " "
            file write meta "z"
            file write meta " "
            file write meta "p"
            file write meta " "
            file write meta  "df" _n

forvalues i=1/`x'{
preserve

keep if gene_num==`i'
metan t ese, nograph randomi
```

```
    file write meta "gene_num"
            file write meta " "
            file write meta "`r(ES)'"
            file write meta " "
            file write meta "`r(seES)'"
            file write meta " "
            file write meta "`r(z)'"
            file write meta " "
            file write meta "`r(p_z)'"
            file write meta " "
            file write meta  "`r(df)'" _n

      restore
      }
file close meta
```

Ανοίγει το αρχείο myfile.txt στο stata και επιλέγονται μόνο τα γονίδια με p-value<0,05

```
import delimited using "myfile.txt"
keep if p<0,05
```

Εντολές διόρθωσης p-value

```
multproc, pval(p)  meth(simes)   rej(simes)

multproc, puncor(0.01) pval(p) meth(simes) rej(fdr)

multproc, pval(p) meth(bonferroni) rej(bonf)

multproc, pval(p) meth(sidak) rej(sidak)

multproc, pval(p) meth(holm) rej(holm)

multproc, pval(p) meth(holland) rej(holland)
```

Λίστα 470 γονιδίων σημαντικών κατά FDR, που τέθηκαν ως είσοδος στο bioCompendium

GeneName	P-value		
mmp25	4.90E-15	Ctsd	3.20E-08
qil1	3.00E-12	trib2	3.30E-08
s100a8	8.70E-12	ust	3.60E-08
ceacam3	1.90E-11	ubl4a	3.80E-08
adrb2	5.40E-11	znf566	4.20E-08
lax1	2.50E-10	alpl	4.60E-08
chst12	2.90E-10	nola2	4.70E-08
Camp	3.20E-10	c18orf17	5.00E-08
loc649986	4.20E-10	klrk1	5.20E-08
Fes	6.20E-10	zyx	5.40E-08
slc19a1	6.40E-10	snapc3	5.50E-08
mrps5	7.10E-10	b3gnt5	5.70E-08
ckap1	8.70E-10	flj10081	5.90E-08
pglyrp1	8.70E-10	znf785	5.90E-08
loc651738	1.10E-09	ptpns1	6.30E-08
rbl1	1.50E-09	znf234	6.50E-08
znf537	1.80E-09	loc283357	7.80E-08
slc16a3	2.20E-09	tmem88	8.00E-08
loc650761	2.50E-09	txndc13	8.10E-08
hk3	2.70E-09	zcchc17	8.90E-08
Hck	3.10E-09	prkcd	1.00E-07
mgc7036	3.10E-09	cryzl1	1.10E-07
dxs9879e	3.40E-09	ibrdc2	1.10E-07
Ncald	3.40E-09	taf15	1.10E-07
plekhg4	3.70E-09	hmfn0839	1.20E-07
atp6v0e2l	4.50E-09	pdcd5	1.30E-07
kiaa0701	4.50E-09	traf3ip3	1.40E-07
eva1	4.80E-09	crygs	1.60E-07
oact1	5.00E-09	hist2h2aa3	1.60E-07
hmg20a	6.10E-09	loc653314	1.60E-07
pafah2	7.60E-09	ppp1r3b	1.60E-07
loc653610	8.50E-09	fprl1	1.70E-07
ccdc76	9.60E-09	sytl2	1.70E-07
st3gal4	1.10E-08	wdr57	1.80E-07
Pdxp	1.40E-08	eif2b2	2.00E-07
anxa3	1.50E-08	tra16	2.00E-07
galnact_2	1.80E-08	galnac4s_6st	2.10E-07
eif3s7	2.20E-08	chd9	2.30E-07
loc652878	2.20E-08	slpi	2.30E-07
Adm	2.30E-08	tkt	2.30E-07
Lat	2.40E-08	ptgdr	2.40E-07
tcn1	2.90E-08	smarcad1	2.50E-07

c3orf9	2.60E-07	tp53inp1	7.80E-07
slamf7	2.60E-07	kiaa0963	8.00E-07
wdr59	2.70E-07	xab1	8.00E-07
gimap6	2.90E-07	rae1	8.20E-07
dkfzp434k1815	3.00E-07	c3orf31	8.50E-07
cecr6	3.10E-07	armc1	8.80E-07
npal3	3.20E-07	phc2	8.90E-07
znf467	3.30E-07	dffb	1.00E-06
fcrl6	3.40E-07	loc349114	1.10E-06
slc6a16	3.40E-07	loc644869	1.10E-06
fez1	3.70E-07	synj2bp	1.10E-06
wdr67	3.80E-07	aplp2	1.20E-06
trappc2	3.90E-07	cbfa2t2	1.20E-06
arg1	4.00E-07	loc645625	1.20E-06
hrihfb2122	4.00E-07	loc647099	1.20E-06
Np	4.10E-07	mefv	1.20E-06
glt1d1	4.50E-07	napepld	1.20E-06
loc400924	4.50E-07	nfe2l2	1.20E-06
rpp21	4.50E-07	sult1b1	1.20E-06
vps45a	4.50E-07	hp	1.30E-06
cxcr7	4.90E-07	kdelc2	1.30E-06
loc440926	4.90E-07	nup43	1.30E-06
nifie14	4.90E-07	dhfrl1	1.40E-06
mosc1	5.00E-07	loc642161	1.40E-06
apobec3g	5.10E-07	loc642755	1.40E-06
ppp4r1	5.60E-07	mgc13170	1.40E-06
sec22l1	5.60E-07	qsox1	1.40E-06
dgat2	5.70E-07	colq	1.50E-06
mgc2463	5.70E-07	loc644128	1.50E-06
mrps10	5.90E-07	notch1	1.50E-06
loc642684	6.10E-07	dffa	1.60E-06
Crot	6.20E-07	gtpbp6	1.60E-06
pxmp4	6.30E-07	nag6	1.60E-06
eif3s6ip	6.40E-07	tgm3	1.60E-06
krt23	6.40E-07	aifm3	1.70E-06
Lbh	6.40E-07	ca4	1.80E-06
tusc4	6.50E-07	hist1h3d	1.80E-06
mms19l	6.60E-07	ikip	1.80E-06
flj10379	6.70E-07	olr1	1.80E-06
rnut1	6.90E-07	arl13b	1.90E-06
coq10a	7.30E-07	flj31413	1.90E-06
c19orf12	7.50E-07	loc339123	1.90E-06
Matk	7.60E-07	ncf1	1.90E-06
rpa1	7.60E-07	cd177	2.00E-06
st3gal6	7.70E-07	flj22662	2.00E-06
steap4	7.70E-07	kiaa1970	2.00E-06

nuak2	2.00E-06	gpr89a	5.60E-06
c9orf111	2.10E-06	loc641825	5.60E-06
flj22471	2.10E-06	loc648716	5.70E-06
hspc176	2.20E-06	loc100190986	5.90E-06
flj20272	2.40E-06	scap1	5.90E-06
sbk1	2.40E-06	bxdc1	6.10E-06
luc7l	2.50E-06	crr9	6.10E-06
ptpn4	2.50E-06	smarcd3	6.10E-06
lsm11	2.60E-06	tnrc5	6.20E-06
znf438	2.60E-06	c1orf183	6.60E-06
gpx3	2.70E-06	gcm1	6.90E-06
gzma	2.70E-06	pbef1	6.90E-06
loc652626	2.70E-06	loc644039	7.10E-06
mgc18216	2.80E-06	serpinb8	7.20E-06
c22orf16	2.90E-06	gls	7.30E-06
statip1	2.90E-06	hyls1	7.40E-06
c4orf29	3.20E-06	erf	7.50E-06
cop1	3.20E-06	ppat	7.60E-06
flvcr2	3.20E-06	hspc196	7.70E-06
eif3s2	3.30E-06	pla2g12a	7.70E-06
nr2c1	3.30E-06	tfb1m	7.70E-06
det1	3.40E-06	arrb2	7.80E-06
grn	3.60E-06	loc284648	8.00E-06
tnfsf13	3.60E-06	atf7ip2	8.20E-06
cited4	3.80E-06	rassf3	8.30E-06
loc440093	3.80E-06	tspan31	8.50E-06
c10orf38	3.90E-06	mett10d	8.70E-06
c6orf166	3.90E-06	myo9a	8.90E-06
dbt	3.90E-06	orf1_fl49	8.90E-06
thoc3	4.00E-06	c17orf60	9.00E-06
hibadh	4.20E-06	heg1	9.10E-06
il8rb	4.20E-06	loc285550	9.10E-06
loc646082	4.20E-06	tm4sf19	9.40E-06
rab43	4.20E-06	kiaa1434	9.60E-06
impdh1	4.30E-06	loc643284	9.70E-06
znf500	4.30E-06	mis12	9.70E-06
c14orf112	4.40E-06	tarp	9.70E-06
kremen1	4.40E-06	c1orf108	1.00E-05
st3gal5	4.70E-06	flot1	0.00001
flj21945	5.10E-06	hoxb2	0.00001
ninj1	5.10E-06	lcmt2	0.00001
znf364	5.10E-06	mgc3121	0.00001
znf519	5.10E-06	mgc4562	0.00001
ephb4	5.20E-06	pogk	1.00E-05
loc652595	5.20E-06	sdccag3	0.00001
impa2	5.30E-06	thap11	0.00001

c9orf19	0.000011	apol3	0.000017
flj10241	0.000011	elovl6	0.000017
klrb1	0.000011	gpr27	0.000017
loc100506828	0.000011	homer3	0.000017
loc284393	0.000011	lrrc39	0.000017
loc652025	0.000011	slc12a6	0.000017
mgc27345	0.000011	cebpa	0.000018
pdcd6ip	0.000011	col18a1	0.000018
serpinb2	0.000011	flj11259	0.000018
tmtc4	0.000011	mgc4093	0.000018
zmym6	0.000011	s100a12	0.000018
c1qr1	0.000012	tlr5	0.000018
casp9	0.000012	agap1	0.000019
il4r	0.000012	agtrap	0.000019
kiaa1026	0.000012	clec5a	0.000019
loc57228	0.000012	dysf	0.000019
scrn1	0.000012	gba	0.000019
spi1	0.000012	loc136143	0.000019
tmco3	0.000012	pycr2	0.000019
csf2rb	0.000013	rab7	0.000019
loc221143	0.000013	slc7a6	0.000019
prf1	0.000013	c21orf127	0.00002
psmb2	0.000013	cpeb2	0.00002
sat	0.000013	cul4a	0.00002
scap2	0.000013	loc441034	0.00002
tsen54	0.000013	c17orf42	0.000021
loc391766	0.000014	ece1	0.000021
ltbr	0.000014	flj33641	0.000021
rab33a	0.000014	mcemp1	0.000021
samd3	0.000014	timp2	0.000021
c9orf5	0.000015	trim68	0.000021
card12	0.000015	fam27e3	0.000022
loc650472	0.000015	ikzf3	0.000022
mgc52498	0.000015	loc653337	0.000022
pbrm1	0.000015	pde6b	0.000022
prok2	0.000015	pscdbp	0.000022
sip1	0.000015	ddx56	0.000023
spin3	0.000015	kiaa0999	0.000023
c15orf39	0.000016	mphosph1	0.000023
dazap2	0.000016	txndc14	0.000023
gfi1	0.000016	zswim4	0.000023
il7r	0.000016	atp6v0b	0.000024
nmt2	0.000016	loc283547	0.000024
plb1	0.000016	phyh2	0.000024
ptch1	0.000016	pnma1	0.000024
abi2	0.000017	dj341d101	0.000025

flj20699	0.000025	c9orf74	0.000035
lin7a	0.000025	kiaa0182	0.000035
echdc3	0.000026	cstf3	0.000037
ptpns1l3	0.000026	flj39501	0.000037
vasp	0.000026	mimitin	0.000037
btbd14a	0.000027	rras2	0.000037
pcnt2	0.000027	slfn13	0.000037
zdhhc18	0.000027	tmem103	0.000037
camta1	0.000028	nfe2	0.000038
loc100499466	0.000028	dnajc24	0.000039
loc100506360	0.000028	loc399900	0.000039
nif3l1	0.000028	tnfaip8l1	0.000039
rab2	0.000028	c7orf25	0.00004
tmem41a	0.000028	gca	0.00004
mnda	0.000029	zmynd11	0.00004
phf15	0.000029	dkfzp564j157	0.000041
rabep1	0.000029	prss15	0.000041
seh1l	0.000029	tspan32	0.000041
slc30a7	0.000029	abcb1	0.000042
timm17b	0.000029	fcgr1a	0.000042
tnfrsf12a	0.000029	loc100287808	0.000042
flj20643	0.00003	plekha1	0.000042
msl3l1	0.00003	aqp9	0.000043
mtp18	0.00003	clec3b	0.000043
npat	0.00003	ffar2	0.000043
basp1	0.000031	loc651143	0.000043
dennd1a	0.000031	rshl2	0.000043
eef2k	0.000031	flj38984	0.000044
nrbp1	0.000031	loc647000	0.000044
rbm28	0.000031	mcee	0.000044
thedc1	0.000031	muted	0.000044
bbs2	0.000032	serpina1	0.000044
bpi	0.000032	klrg1	0.000045
c16orf24	0.000032	limk2	0.000045
flj10769	0.000032	mll	0.000045
pcsk7	0.000032	tmem117	0.000045
adora3	0.000033	akap11	0.000046
dpep3	0.000033	c1orf33	0.000046
hspc117	0.000033	dclre1a	0.000046
map1lc3b2	0.000033	fam20b	0.000046
flj20551	0.000034	flj21749	0.000046
loc652615	0.000034	tbc1d2	0.000046
mga	0.000034	zcsl2	0.000046
pgd	0.000034	loc285053	0.000047
rab11fip3	0.000034	ncrna00182	0.000047
znf562	0.000034	vezt	0.000048

znf545	0.000048
mgc40499	0.000049
tp53rk	0.000049
plekha9	0.00005
sirpd	0.00005
dyrk2	0.000051
phc1	0.000053
rnase4	0.000054
orm1	0.000055
gpd1l	0.000056
loc152485	0.000057
trim73	0.000057
loc654053	0.000058
oplah	0.000058
m_rip	0.000059
jtv1	0.00006
kif13a	0.00006
lima1	0.00006