



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ
ΒΙΟΙΑΤΡΙΚΗ»**

**Μεθοδολογίες Υπολογιστικής Ανάλυσης και Μέτα-ανάλυσης
γενετικών δεδομένων και δεδομένων γονιδιακής έκφρασης,
εφαρμογή στο Σακχαρώδη Διαβήτη**

Παπαδημητρίου Κωνσταντίνα

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
ΜΠΑΓΚΟΣ ΠΑΝΤΕΛΗΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ**

Λαμία, 2016



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ
ΚΑΤΕΥΘΥΝΣΗ «ΥΠΟΛΟΓΙΣΤΙΚΗ ΙΑΤΡΙΚΗ ΚΑΙ
ΒΙΟΛΟΓΙΑ»**

**Μεθοδολογίες Υπολογιστικής Ανάλυσης και Μέτα-ανάλυσης
γενετικών δεδομένων και δεδομένων γονιδιακής έκφρασης,
εφαρμογή στο Σακχαρώδη Διαβήτη**

Παπαδημητρίου Κωνσταντίνα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΜΠΑΓΚΟΣ ΠΑΝΤΕΛΗΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ**

Λαμία, 2016

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο [«τίτλος εργασίας»] αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Η ΔΗΛΟΥΣΑ
ΠΑΠΑΔΗΜΗΤΡΙΟΥ ΚΩΝΣΤΑΝΤΙΝΑ

Ημερομηνία

Υπογραφή

**Μεθοδολογίες Υπολογιστικής Ανάλυσης και Μέτα-ανάλυσης
γενετικών δεδομένων και δεδομένων γονιδιακής έκφρασης,
εφαρμογή στο Σακχαρώδη Διαβήτη**

Παπαδημητρίου Κωνσταντίνα

Τριμελής Επιτροπή:

ΜΠΑΓΚΟΣ ΠΑΝΤΕΛΗΣ (επιβλέπων)

ΠΛΑΓΙΑΝΝΑΚΟΣ ΒΑΣΙΛΕΙΟΣ

ΑΛΑΜ ΜΑΡΙΑ

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική κ.Μπάγκο Παντελή για την ευκαιρία που μου έδωσε να εκπονήσω το συγκεκριμένο κομμάτι έρευνας υπό την επίβλεψή του, την εμπιστοσύνη που μου επέδειξε, τη συνεχόμενη καθοδήγηση και ώθηση αλλά και τη πολύτιμη βοήθεια του γενικότερα στην διάρκεια εκπόνησης της εργασίας μου.

Επίσης θα ήθελα να εκφράσω ιδιαίτερα τις ευχαριστίες στην υποψήφια διδάκτορα Κοντού Παναγιώτα για την καθοδήγηση και τον έλεγχο των πειραματικών αποτελεσμάτων σε όλα τα στάδια εξέλιξης της εργασίας μέχρι και την ολοκλήρωσή της.

Τέλος θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα Βέννου Κωνσταντίνα για την εξαιρετική συνεργασία μας και για το πνεύμα ομαδικότητας που την διακρίνει.

Περίληψη:

Στη παρούσα διπλωματική εργασία, πραγματοποιήθηκε μετα-ανάλυση γονιδίων, με τη χρήση του στατιστικού πακέτου STATA13, ξεχωριστά για κάθε τύπο σακχαρώδη διαβήτη προκειμένου να εντοπιστούν τα γονίδια τα οποία εκφράζονται διαφορεικά στη νόσο (του σακχαρώδη διαβήτη).

Κατά τη περάτωση της μετα-ανάλυσης προέκυψαν τα στατιστικώς σημαντικά γονίδια στα οποία εφαρμόστηκαν μέθοδοι διόρθωσης των p-value προκειμένου να ενισχυθεί το εύρημα. Ωστόσο κατά την εφαρμογή των μεθόδων δε βρέθηκε σημαντική συσχέτιση κάποιου γονιδίου με την ασθένεια. Παρόλο αυτά, η συμμετοχή σε μεταβολικές δραστηριότητες προσδίνουν την ύπαρξη προδιάθεσης για εμπλοκή στη νόσο. Τα εργαλεία της BioCompendium, Panther και της STRING χρησιμοποιήθηκαν για να συλληφθούν τα μοριακά και βιοχημικά χαρακτηριστικά των επιλεγμένων γονιδίων αλλά και οι τυχόν αλληλεπιδράσεις μεταξύ αυτών.

Αν και τα αποτελέσματα δίνουν μία «κατεύθυνση», δεν είναι ισχυρά ώστε να ισχυριστούμε την συσχέτιση των γονιδίων αυτών με τη νόσο. Χρειάζεται περαιτέρω ανάλυση αυτών για να τεκμηριωθεί και να αξιολογηθεί η συσχέτιση τόσο με την εκδήλωση του σακχαρώδη διαβήτη όσο και με άλλες μεταβολικές ασθένειες.

Λέξεις κλειδιά: Μικροσυστοιχίες, Σακχαρώδης Διαβήτης, μετα-ανάλυση, διαφορική έκφραση γονιδίων, bioCompendium, Panther, STRING

Abstract:

For the purpose of this master thesis, a gene metanalysis has be done with the statistical software package STATA 13. The metanalysis held for every single one type of diabetes mellitus in order to locate the genes which are differentially expressed at the disease.

At the whole phase of the metanalysis statistically important genes were revealed. In these results a p-value statistical correction method applied in order to strengthened the findings of the results. However, when these methods applied the findings had not a significant relation between the disease and these genes. Nevertheless, participation in metabolic activities confer a predisposition for being involved in the disease. The BioCompendium , Panther and STRING tools used to capture the molecular and biochemical characteristics of the selected genes plus the possible interaction between them.

Although the results give a "direction" but are not strong enough to assert the relevance of these genes to disease. Further analysis to document and to assess the association with both the occurrence of diabetes and other metabolic diseases.

Keywords: Microarrays, Diabetes Mellitus, meta-analysis, differentially expressed genes, bioCompendium, Panther, STRING

Περιεχόμενα

Εισαγωγή:	2
1 Κεφάλαιο:	4
1.1 Βασικές Έννοιες:	4
1.2 Σακχαρώδης διαβήτης:	5
1.2.1 Τύποι Σακχαρώδη διαβήτη:	5
1.2.2 Η κατάσταση στην Ελλάδα:	6
1.3 Μικροσυστοιχίες:	8
1.4 Βασικά βήματα για ένα πείραμα μικροσυστοιχιών:	10
1.4.1 Διατύπωση βιολογικής ερώτησης:	10
1.4.2 Επιλογή του κατάλληλου τύπου μικροσυστοιχίας:	11
1.4.3 Παρασκευή δείγματος:	12
1.4.4 Υβριδοποίηση:	12
1.4.5 Σάρωση:	13
1.5 Ποσοτικοποίηση των δεδομένων:	14
1.6 Κανονικοποίηση:	15
1.6.1 Κανονικοποίηση ολικής έντασης:	15
1.6.2 Lowess (γραμμική οπισθοδρόμηση τοπικών βαρών) κανονικοποίηση:	16
2 Κεφάλαιο:	17
2.1 Ανάλυση μικροσυστοιχιών-Στατιστική ανάλυση:	17
2.1.1 T-test:	17
2.1.2 Μέθοδος Permutation:	17
2.1.3 Μέθοδος Bootstrap:	18
2.1.4 Μέθοδοι διόρθωσης του p-value:	19
2.2 Ανάλυση μικροσυστοιχιών-Ομαδοποίηση(Clustering):	21
2.3 Ανάλυση μικροσυστοιχιών-Πρόγνωση:	22
2.3.1 Ο αλγόριθμος ιεραρχικής ταξινόμησης:	22
2.3.2 Οι διαχωριστικοί αλγόριθμοι ομαδοποίησης:	23
2.3.3 Μέθοδοι επιβλεπόμενης μάθησης:	23
2.3.4 Μέθοδοι μη επιβλεπόμενης μάθησης:	24
2.4 Βάσεις δεδομένων γονιδιακής έκφρασης:	24
2.5 Μετα-Ανάλυση (Meta-Analysis):	25
2.5.1 Το μοντέλο σταθερών επιδράσεων (fixed effect model):	25
2.5.2 Το μοντέλο τυχαίων επιδράσεων (random effect model):	26
2.5.3 Προβλήματα κατά τη διεξαγωγή μίας μετα-ανάλυσης:	27

2.6	Η εφαρμογή της μετα-ανάλυσης στις μικροσυτοιχίες (Meta-analysis in Microarrays):	27
2.6.1	Μέθοδος των μεγεθών επίδρασης:	28
2.6.2	Μέθοδος συνδυασμού των p-values:	28
2.6.3	Μέθοδος υπολογισμού του γινομένου των βαθμών κατάταξης (rank product): 28	
3	Κεφάλαιο -Υλικά και Μέθοδοι:	29
3.1	Ερευνητικό ερώτημα:	29
3.2	Συλλογή και καταγραφή των δεδομένων:	29
3.3	Στατιστική ανάλυση των δεδομένων:	31
3.4	Η χρήση των πλατφορμών bioCompedium, Panther και STRING :	33
4	Κεφάλαιο -Αποτελέσματα και Συζήτηση:	35
4.1	Ανάκτηση δεδομένων από τη βάση GEO:	35
4.2	Δεδομένα προς στατιστική ανάλυση και μετα-ανάλυση :	36
4.3	Εύρεση των στατιστικώς σημαντικών γονιδίων:	37
4.4	Τα δεδομένα ως είσοδο στις πλατφόρμες bioCompedium, Panther και STRING:	42
4.4.1	Για το σακχαρώδη διαβήτη τύπου 1:	43
4.4.2	Για το σακχαρώδη διαβήτη τύπου 2:	49
4.5	Σύνοψη:	53
	Βιβλιογραφία:	55
	Παράρτημα 1:	57
	Παράρτημα 2:	59

Εισαγωγή:

Η παρούσα διπλωματική εργασία, εκπονήθηκε στα πλαίσια του μεταπτυχιακού μου προγράμματος «Πληροφορικής και Υπολογιστικής Βιοϊατρικής», του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας και έλαβε χώρα στη Λαμία. Ο τίτλος της εργασίας είναι «Μεθοδολογίες Υπολογιστικής Ανάλυσης και Μετα-ανάλυσης γενετικών δεδομένων και δεδομένων γονιδιακής έκφρασης, εφαρμογή στο Σακχαρώδη διαβήτη» και υλοποιήθηκε από τον Φεβρουάριο 2016 έως και τον Οκτώβριο 2016.

Το κεντρικό θέμα της παρούσας διπλωματικής είναι η απάντηση του παρακάτω ερωτήματος «Ποιες είναι οι διαφορές στις τιμές της γονιδιακής έκφρασης ανάμεσα σε κύτταρα υγιών και ασθενών που πάσχουν από σακχαρώδη διαβήτη». Σκοπός είναι να προσδιοριστούν γονίδια τα οποία συσχετίζονται με την εμφάνιση της νόσου.

Προκειμένου να απαντηθεί το ερώτημά μας, πραγματοποιήθηκε μετα-ανάλυση μικροσυστοιχιών. Η συγκεκριμένη μεθοδολογία εφαρμόζεται ευρέως από επιστήμονες που διερευνούν τέτοιας φύσεως προβλήματα όπως είναι για παράδειγμα οι διαφορές στις τιμές γονιδιακής έκφρασης κατά τη χορήγηση φαρμάκου ή η εφαρμογή θεραπείας.

Υπάρχουν δημοσιεύσεις στη βιβλιογραφία οι οποίες εφαρμόζουν τη μέθοδο της μετα-ανάλυσης προκειμένου να δώσουν απαντήσεις στα διάφορα ερωτήματα. Ωστόσο η εφαρμογή της μετα-ανάλυσης σε μικροσυστοιχίες είναι περιορισμένες. Ενδεικτικά μερικές από αυτές είναι η μελέτη [Daniel R. Rhodes et al, 2002] στην οποία αποκαλύφθηκε βιοχημικό μονοπάτι το οποίο συνδέεται με τη δυσλειτουργία του καρκίνου του προστάτη, η μελέτη [Jonathan A. Ewald et al, 2013] στην οποία αναλύθηκαν όγκοι της ουροδόχου κύστης σε διάφορα στάδια προκειμένου να ανιχνευτεί η αυξημένη έκφραση των Ras/MAPK και PI3K ως πιθανά ρυθμιστικά σήματα κατατεθέν (βιοδείκτες) της εξέλιξης της νόσου, η μελέτη [Anders M. et al, 2013] στην οποία καθορίστηκαν οι γενετικές υπογραφές που σχετίζονται με την αγγειογένεση στα ανθρώπινα κύτταρα, η μελέτη [Burguillo et al, 2010] όπου γίνεται εξέταση της αντίστασης του imatinib (φάρμακο που χορηγείται σε πολλές μορφές καρκίνου) στη χρόνια μυελογενή λευχαιμία.

Η διεξαγωγή της συγκεκριμένης μεθοδολογίας είναι μία αρκετά σύνθετη διαδικασία. Απαιτείται από το χρήστη της να κατέχει ένα καλό υπόβαθρο γνώσεων τόσο στο υπολογιστικό όσο και στο βιολογικό κομμάτι. Όπως είναι αναμενόμενο, η ακρίβεια, η συνέπεια και η τεράστια προσοχή παίζουν καθοριστικό ρόλο στη σωστή διεξαγωγή της και ανάλογα με τη διαθεσιμότητα των δεδομένων μπορεί να καταστεί χρονοβόρα.

Για να γίνει αντιληπτή με μία πρώτη ματιά η διαδικασία για την οποία γίνεται λόγος στη συγκεκριμένη διπλωματική αναλύονται παρακάτω επιγραμματικά τα κεφάλαια που την αποτελούν. Η εργασία αποτελείται από 4 κύρια κεφάλαια.

Στο πρώτο κεφάλαιο, αρχικά γίνεται λόγος για τις βασικές έννοιες της επιστήμης της βιοπληροφορικής. Κατόπιν δίνεται ο ορισμός αλλά και η διεξοδική ανάλυση ενός πειράματος μικροσυστοιχιών. Τέλος, γίνεται αναφορά στη μεταβολική νόσο του σακχαρώδη διαβήτη, στους διαφόρους τύπους της και στα επίσημα στοιχεία που Παγκόσμιος Οργανισμός Υγείας (Π.Ο.Υ.) για την κατάσταση της νόσου που επικρατεί στην Ελλάδα.

Στο δεύτερο κεφάλαιο, παρουσιάζονται οι στατιστικές μέθοδοι που χρησιμοποιήθηκαν μεμονωμένα για κάθε μελέτη (t-test, fold change, permutation test, bootstrap) όπως επίσης και των στατιστικών μεθόδων που χρησιμοποιήθηκαν στη μετα-ανάλυση (μέθοδος του ανασυνδυασμού των p-values, των μεγεθών επίδρασης και του γινομένου των βαθμών κατάταξης). Επίσης αναφέρονται οι μέθοδοι διόρθωσης του p-value

για την αποφυγή λανθασμένων στατιστικά σημαντικών γονιδίων. Τέλος δίνεται ο ορισμός της μετα-ανάλυσης και της προσφοράς της στις μικροσυστοιχίες.

Στο τρίτο κεφάλαιο παρουσιάζονται τα υλικά και οι μέθοδοι που χρησιμοποιήθηκαν για τη διεξαγωγή της εν λόγω διπλωματικής. Πιο συγκεκριμένα, παρουσιάζεται η μεθοδολογία η οποία ακολουθήθηκε προκειμένου να εξαχθούν τα αποτελέσματα, τα οποία δίνουν την απάντηση στο ερώτημά που τέθηκε. Πολύ επιγραμματικά, αρχικά καθορίζεται το ερώτημα, έπειτα πραγματοποιείται η αναζήτηση και καταγραφή των δεδομένων από τη βάση δεδομένων της GEO (Gene Expression Omnibus), επεξεργασία αυτών των δεδομένων, με έλεγχο t-test & bootstrap (στη περίπτωση που ελέγχεται κάθε μελέτη χωριστά) και με μετα-ανάλυση (στη περίπτωση της σύνθεσης των μελετών). Εκτελούνται οι μέθοδοι διόρθωσης του p-value προκειμένου να αποφευχθούν λανθασμένα θετικά στατιστικά σημαντικά γονίδια (false positive). Γίνεται εντοπισμός των στατιστικά σημαντικών γονιδίων και με τη χρήση διαφόρων διαδικτυακών πλατφορμών (bioCompendium, Panther, STRING) επιστρέφονται πληροφορίες των γονιδίων αυτών σχετικά με τις βιολογικές, τις μοριακές λειτουργίες, τα βιοχημικά-μεταβολικά μονοπάτια στα οποία μπορεί να εμπλέκονται καθώς και τα δίκτυα αλληλεπιδράσεων μεταξύ αυτών.

Στο τέταρτο κεφάλαιο δίνεται βάρος στα αποτελέσματα της μεθοδολογίας που αναλύθηκε στο προηγούμενο (κεφάλαιο) αλλά και της ερμηνείας αυτών. Αναλυτικότερα, παρουσιάζεται η επιστροφή που δόθηκε μετά την εφαρμογή του ερωτήματος στη βάση GEO. Δίνεται ο τελικός αριθμός των μελετών με τα διάφορα χαρακτηριστικά τους, οι οποίες χρησιμοποιήθηκαν στη παρούσα εργασία. Η επεξεργασία των δεδομένων που αντλήθηκαν με τέτοιο τρόπο ώστε να μπορούν να κατευθυνθούν προς τη διαδικασία της μετα-ανάλυσης. Τα αποτελέσματα της μετα-ανάλυσης που ήταν τα στατιστικώς σημαντικά γονίδια. Το κεφάλαιο ολοκληρώνεται με την παράθεση και το σχολιασμό των αποτελεσμάτων.

Ακολουθεί η Βιβλιογραφία και τα Παραρτήματα 1 και 2. Στο Παράρτημα 1 της εργασίας ο αναγνώστης μπορεί να βρει τις μελέτες (GSEs) οι οποίες απορριφθήκαν και στο Παράρτημα 2 παραθέτεται ο κώδικας που χρησιμοποιήθηκε στο πρόγραμμα του STATA13.

1 Κεφάλαιο:

1.1 Βασικές Έννοιες:

Στη σημερινή εποχή ο τεράστιος όγκος πληροφορίας που προκύπτει από την έρευνα στις Βιολογικές επιστήμες, κυρίως για μόρια όπως το DNA, το RNA και οι αλληλουχίες των πρωτεϊνών, απαιτεί εξαιρετικά εξελιγμένα υπολογιστικά συστήματα προκειμένου τα δεδομένα να καταχωρηθούν και να ταξινομηθούν, ώστε να καταστούν “human friendly” και να αξιοποιηθούν από τους ερευνητές.

Η Βιοπληροφορική, αποτελεί έναν ολοκληρωμένο επιστημονικό κλάδο που αναπτύχθηκε και διεκδίκησε χώρο στο ενδιάμεσο της Βιολογίας και της Πληροφορικής, όταν πια τα βιολογικά δεδομένα ήταν πολλά και δύσκολα επεξεργάσιμα από τον ανθρώπινο παράγοντα, αλλά και όταν η Πληροφορική είχε φτάσει σε επαρκές επίπεδο να προσφέρει βοήθεια [N. M. Luscombe, 2001].

ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ:

«Η επιστήμη που εφαρμόζει υπολογιστικές μεθόδους (επιστήμη υπολογιστών, εφαρμοσμένα μαθηματικά, στατιστική) με σκοπό την οργάνωση, διαχείριση και κατανόηση της πληροφορίας που σχετίζεται με βιομακρομόρια (DNA, RNA, πρωτεΐνες, πολυσακχαρίτες)»

Κύριοι στόχοι της Βιοπληροφορικής:

- ✓ Αποδοτική οργάνωση των βιολογικών δεδομένων και πρόσβαση σε αυτά, καθώς και συσσώρευση νέων δεδομένων.
- ✓ Ανάπτυξη μεθόδων και υπολογιστικών εργαλείων με στόχο την εξαγωγή πληροφοριών από τα δεδομένα.
- ✓ Χρήση των εργαλείων αυτών για την ανάλυση και ερμηνεία των δεδομένων με ένα βιολογικά αποδεκτό τρόπο.

Το πρόβλημα που υπήρχε ήταν στην πολυπαραγοντικότητα των δεδομένων, λόγω του μεγάλου εύρους τους. Απαιτούνται εργαλεία - προγράμματα ανάλυσης των δεδομένων που θα επιτρέπουν την πολύ-παραμετρική στατιστική ανάλυση και την κατανοητή απεικόνιση των αποτελεσμάτων. Έτσι αναπτύχθηκαν και αναπτύσσονται αλγόριθμοι ανάλυσης και κατασκευάζονται προγράμματα ανάλυσης φιλικά προς τον χρήστη - ερευνητή.

Μία από αυτές τις τεχνολογίες που έδωσαν λύση στην επεξεργασία και ανάλυση χιλιάδων δεδομένων ήταν οι Μικροσυστοιχίες. Μπορούν να υπολογίσουν ταυτόχρονα την έκφραση χιλιάδων γονιδίων σε διαφορετικά δείγματα ή σε διαφορετικά στάδια ανάπτυξης, χρησιμοποιούνται για τη κατανόηση βιολογικών μηχανισμών (παρέχουν χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια ενεργοποιούνται ή καταστέλλονται σε διάφορα στάδια ανάπτυξης ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία την εύρεση νέων γονιδίων τα οποία σχετίζονται με ασθένειες), την απόκριση σε φαρμακευτικές ουσίες ή θεραπείες και τη σύγκριση έκφρασης σε φυσιολογικές και παθολογικές καταστάσεις (Εικόνα 1).



Εικόνα 1 : Gene expression profiling

1.2 Σακχαρώδης διαβήτης:

Ο Σακχαρώδης διαβήτης είναι μία σοβαρή χρόνια ασθένεια η οποία εμφανίζεται είτε όταν το πάγκρεας δεν παράγει αρκετή ινσουλίνη (μία ορμόνη που ελέγχει το σάκχαρο στο αίμα) είτε όταν ο οργανισμός δεν μπορεί να χρησιμοποιήσει αποτελεσματικά την ινσουλίνη που παράγει. Ο διαβήτης αποτελεί ένα σημαντικό πρόβλημα δημόσιας υγείας, τόσο ο αριθμός των περιπτώσεων όσο και ο επιπολασμός του σακχαρώδη διαβήτη έχουν αυξηθεί σταθερά κατά τη διάρκεια των τελευταίων δεκαετιών.

Παγκοσμίως, έχει υπολογισθεί ότι μέχρι το 2014, 422 εκατομμύρια ενήλικες έπασχαν από σακχαρώδη διαβήτη, σε σύγκριση με τα 108 εκατομμύρια ενηλίκων του 1980. Η παγκόσμια επικράτηση του σακχαρώδη διαβήτη έχει σχεδόν διπλασιαστεί από το 1980, αυξήθηκε από 4,7% σε 8,5% του ενήλικου πληθυσμού. Η εξέλιξη αυτή αντανακλά την αύξηση των σχετιζόμενων παραγόντων κινδύνου, όπως είναι παχυσαρκία. Κατά την τελευταία δεκαετία, ο επιπολασμός του διαβήτη έχει αυξηθεί ταχύτερα σε χαμηλού και μεσαίου εισοδήματος χώρες από ό, τι στις χώρες υψηλού εισοδήματος [(WHO), 2016].

1.2.1 Τύποι Σακχαρώδη διαβήτη:

Ο διαβήτης όλων των τύπων μπορεί να οδηγήσει σε επιπλοκές σε πολλά μέρη του σώματος και μπορεί να αυξήσει τον συνολικό κίνδυνο πρόωρου θανάτου. Πιθανές επιπλοκές μπορεί να είναι: καρδιακή προσβολή, εγκεφαλικό επεισόδιο, νεφρική ανεπάρκεια, ακρωτηριασμός άκρων, απώλεια όρασης και νευρική βλάβη. Στην εγκυμοσύνη, αυξάνει τον κίνδυνο θανάτου του εμβρύου και άλλες επιπλοκές. Οι κύριοι τύποι σακχαρώδους διαβήτη είναι: ο σακχαρώδης διαβήτης τύπου 1, τύπου 2 και ο διαβήτης κύησης. Πιο συγκεκριμένα:

Ο διαβήτης τύπου 1 (παλαιότερα γνωστός ως ινσουλινο-εξαρτώμενος, νεανικός ή παιδικός διαβήτης) χαρακτηρίζεται από ανεπαρκή παραγωγή ινσουλίνης στο σώμα. Τα άτομα με διαβήτη τύπου 1 απαιτούν καθημερινή χορήγηση ινσουλίνης για τη ρύθμιση της ποσότητας της γλυκόζης στο αίμα τους. Εάν δεν έχουν πρόσβαση στην ινσουλίνη, δεν μπορούν να επιβιώσουν. Η αιτία του διαβήτη τύπου 1 δεν είναι γνωστή και επί του παρόντος δεν μπορεί να προληφθεί. Κύρια συμπτώματά του είναι: η υπερβολική ούρηση και δίψα, συνεχή πείνα, απώλεια βάρους, αλλαγές στην όραση και κόπωση.

Ο διαβήτης τύπου 2 (παλαιότερα γνωστός ως μη-ινσουλινεξαρτώμενος ή διαβήτης ενηλίκων) χαρακτηρίζεται από την αναποτελεσματική χρήση της ινσουλίνης. Ο διαβήτης τύπου 2 αποτελεί τη συντριπτική πλειοψηφία των διαβητικών παγκοσμίως. Τα συμπτώματα μπορεί να είναι παρόμοια με εκείνα του διαβήτη τύπου 1, αλλά συχνά είναι λιγότερο έντονα ή ακόμα και να απουσιάζουν. Έτσι σαν αποτέλεσμα έχει τη μη έγκαιρη διάγνωση μέχρι ως ότου τελικά προκύψουν επιπλοκές. Για πολλά χρόνια ο διαβήτης τύπου 2 παρατηρήθηκε μόνο σε ενήλικες, αλλά έχει αρχίσει να εμφανίζεται και σε παιδιά.

Ο διαβήτης κύησης είναι μια προσωρινή κατάσταση που εμφανίζεται στην εγκυμοσύνη και φέρει μακροπρόθεσμο κίνδυνο για εμφάνιση σακχαρώδους διαβήτη τύπου 2. Η κατάσταση είναι παρούσα όταν οι τιμές της γλυκόζης στο αίμα είναι πάνω από το φυσιολογικό, αλλά κάτω από τα διαγνωστικά του σακχαρώδη διαβήτη. Οι γυναίκες με διαβήτη κύησης διατρέχουν αυξημένο κίνδυνο κάποιων επιπλοκών κατά τη διάρκεια της εγκυμοσύνης και του τοκετού, όπως επίσης και τα βρέφη τους. Ο διαβήτης κύησης μπορεί να διαγνωστεί με προγεννητικό έλεγχο. [(WHO), 2016]

1.2.2 Η κατάσταση στην Ελλάδα:

Προκειμένου να αντιμετωπισθεί η παγκόσμια εξάπλωση του διαβήτη και των μη μεταδοτικών ασθενειών (noncommunicable diseases-NCDs), έγινε επιτακτική η ανάγκη να δημιουργηθεί μια βάση για την παρακολούθηση των τάσεων και να αξιολογηθεί η πρόοδος των χωρών στην αντιμετώπιση της επιδημίας. Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ), δημιούργησε «προφίλ» του διαβήτη για κάθε χώρα. Κάθε προφίλ περιλαμβάνει στοιχεία για τον εξάπλωση και τις τάσεις του διαβήτη, τη θνησιμότητα, τους παράγοντες κινδύνου, την παρακολούθηση και την επιτήρηση, τις πολιτικές πρωτογενούς πρόληψης και θεραπείας, τη διαθεσιμότητα των φαρμάκων, τη διαθεσιμότητα βασικών τεχνολογιών και διαδικασιών [(WHO), 2016].

Σύμφωνα με τα επίσημα στοιχεία που παραθέτει ο Παγκόσμιος Οργανισμός Υγείας για την Ελλάδα, η θνησιμότητα για τις ηλικίες 30-69 είναι μεγαλύτερη για τους άνδρες, ενώ από την ηλικία των 70 και πάνω η θνησιμότητα μεγαλώνει για τις γυναίκες και μειώνεται για τους άνδρες. Ωστόσο με τη πάροδο των ετών (σε κανονικοποιημένη κλίμακα), το ποσοστό των ανδρών που πάσχουν από σακχαρώδη διαβήτη αυξάνεται σε σχέση με το ποσοστό των γυναικών. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι το ποσοστό θνησιμότητας για όλες τις ηλικίες των διαβητικών καταλαμβάνει μόλις το 1% (Εικόνα 2).

Όπως ήδη έχει αναφερθεί κύριοι παράγοντες εξάπλωσης του διαβήτη είναι: η παχυσαρκία, η φυσική αδράνεια και η συνεχόμενη πρόσληψη βάρους (Εικόνα 3).

Greece

Total population: 10 955 000

Income group: High

Mortality*

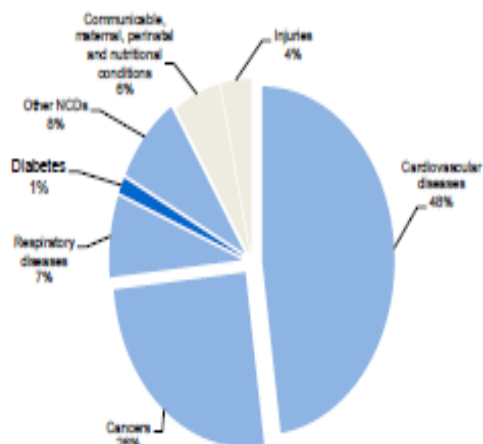
Number of diabetes deaths

	males	females
ages 30–69	190	<100
ages 70+	570	730

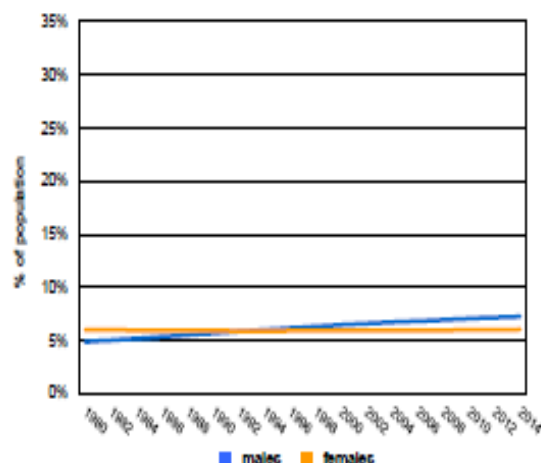
Number of deaths attributable to high blood glucose

	males	females
ages 30–69	780	280
ages 70+	2 290	3 870

Proportional mortality (% of total deaths, all ages)*



Trends in age-standardized prevalence of diabetes



Εικόνα 2: Κατάσταση θνησιμότητας στην Ελλάδα, πηγή:WHO

Prevalence of diabetes and related risk factors

	males	females	total
Diabetes	9.5%	8.8%	9.1%
Overweight	69.6%	60.2%	64.9%
Obesity	23.6%	26.7%	25.1%
Physical inactivity	12.4%	18.2%	15.4%

Εικόνα 3:Επιπολασμός διαβήτη και οι παράγοντες κινδύνου, πηγή:WHO

National response to diabetes

Policies, guidelines and monitoring

Operational policy/strategy/action plan for diabetes	Yes†
Operational policy/strategy/action plan to reduce overweight and obesity	Yes†
Operational policy/strategy/action plan to reduce physical inactivity	Yes
Evidence-based national diabetes guidelines/protocols/standards	Not available
Standard criteria for referral of patients from primary care to higher level of care	Not available
Diabetes registry	No
Recent national risk factor survey in which blood glucose was measured	No

Availability of medicines, basic technologies and procedures in the public health sector

Medicines in primary care facilities

Insulin	●
Metformin	●
Sulphonylurea	●

Procedures

Retinal photocoagulation	●
Renal replacement therapy by dialysis	●
Renal replacement therapy by transplantation	●

Basic technologies in primary care facilities

Blood glucose measurement	●
Oral glucose tolerance test	●
HbA1c test	●
Dilated fundus examination	○
Foot vibration perception by tuning fork	●
Foot vascular status by Doppler	○
Urine strips for glucose and ketone measurement	●

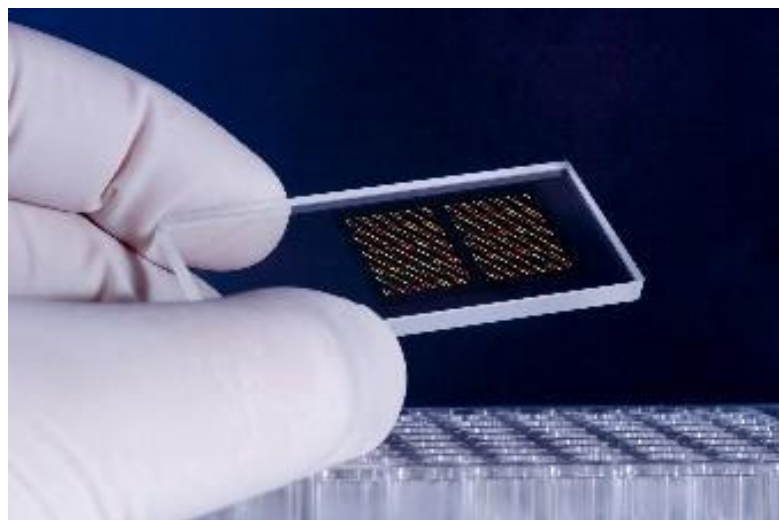
Εικόνα 4:Εθνική αντιμετώπιση του διαβήτη, πηγή:WHO

1.3 Μικροσυστοιχίες:

ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ:

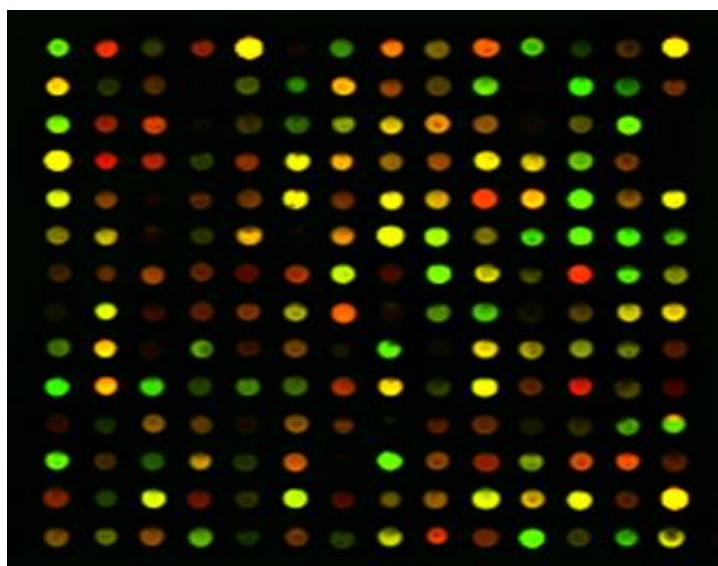
Γυάλινο πλακίδιο που αποτελείται από συγκεκριμένες αλληλουχίες οι οποίες είναι ειδικές για συγκεκριμένα γονίδια, τους ανιχνευτές (probes), οι οποίοι είναι ακινητοποιημένοι σε μία κουκκίδα (spot) της γυάλινης επιφάνειας του πλακιδίου (Εικόνα 5) [Lee, 2004].

Οι μικροσυστοιχίες παρέχουν τη δυνατότητα απεικόνισης της έκφρασης χιλιάδων ή δεκάδων χιλιάδων γονιδίων σε δεκάδες ή εκατοντάδες δείγματα. Οι διατάξεις αυτές μπορούν να υποστούν ταξινόμηση, ομαδοποίηση, εκτίμηση πυκνότητας. Για παράδειγμα, μετρώντας επίπεδα έκφρασης σχετιζόμενα με δύο τύπους ιστών (φυσιολογικός ή ασθενής) δημιουργείται ένα σύνολο δεδομένων με ετικέτες που μπορεί να χρησιμοποιηθεί για διαγνωστική ταξινόμηση. Αναλυτικότερα, τα γονίδια συνήθως βρίσκονται σε δύο διακριτές βιολογικές καταστάσεις: καταστολή (repression, off) ή ενεργοποίηση (induction, on). Έτσι, κάθε γονίδιο παρουσιάζει διαφορετικά επίπεδα έκφρασης κατά μήκος των δειγμάτων που δεν ανήκουν στο ίδιο είδος ιστών (φυσιολογικός ή ασθενής) [Eric P. Xing, 2001].



Εικόνα 5: Chip μικροσυστοιχιών

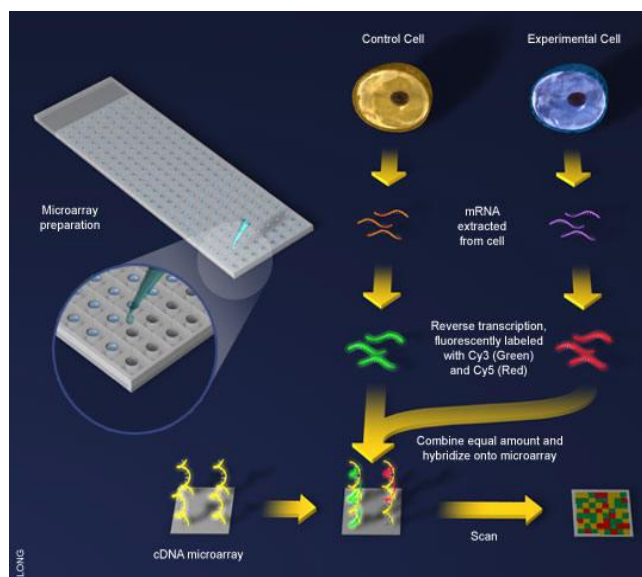
Πιο συγκεκριμένα, πολλά ή όλα τα γονίδια ενός οργανισμού αντιπροσωπεύονται από νουκλεοτιδικές αλληλουχίες (ανιχνευτές) ανά σημείο σε μεγάλη πυκνότητα στο γυάλινο πλακάκι της μικροσυστοιχίας. Το mRNA το οποίο έχει απομονωθεί από 2 τύπους κυττάρων ή ιστών (φυσιολογικός-ασθενής), μετατρέπεται σε DNA και σημαίνεται με φθορίζουσα ουσία. Το σημασμένο δείγμα του DNA, το οποίο έχει προκύψει από το mRNA, υβριδοποιείται στο πλακάκι και σαρώνεται με σαρωτή, για να μετρηθεί η ένταση του σήματος κάθε σημείου. Το αποτέλεσμα της σάρωσης είναι μία εικόνα με διάφορες διαβαθμίσεις χρωμάτων (Εικόνα 6). Η τιμή της έντασης σήματος θεωρείται ότι είναι αντιπροσωπευτική της ποσότητας του μεταγράφου κάθε γονιδίου στο δείγμα. Στη συνέχεια, τα δεδομένα ανάλυσης εξάγουν συμπεράσματα για το ποια γονίδια εμφανίζουν διαφοροποιήσεις στην ένταση [D. Stekel, 2003].



Εικόνα 6: Συνδυασμένη εικόνα που προκύπτει μετά τη σάρωση των μικροσυστοιχιών.

1.4 Βασικά βήματα για ένα πείραμα μικροσυστοιχιών:

Για να εκπληρωθεί με επιτυχία ένα πείραμα μικροσυστοιχιών, απαιτείται μία σειρά από ακριβείς διεργασίες. Μία αναπαράσταση των σταδίων που ακολουθούνται παρουσιάζεται στην Εικόνα 7. Όπως είναι φανερό η όλη διαδικασία είναι αρκετά περίπλοκη και για να στεφθεί με επιτυχία το πείραμά μας πρέπει να εξασφαλιστεί η αποδεκτή ποιότητα των δεδομένων που θα προκύψουν και των αποτελεσμάτων που θα εξαχθούν.

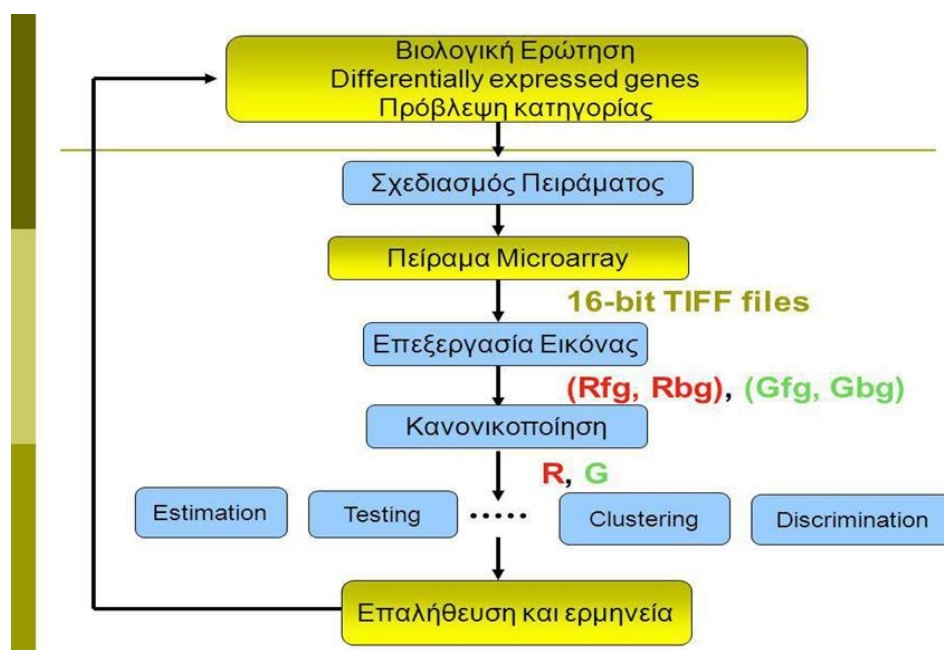


Εικόνα 7: Διαδικασία διεξαγωγής πειράματος μικροσυστοιχιών

1.4.1 Διατύπωση βιολογικής ερώτησης:

Αρχικά, για να καταστεί δυνατή η διαδικασία πειράματος μικροσυστοιχιών είναι αναγκαίο να διατυπωθεί με ακρίβεια το ερώτημά μας. Αυτό θα μπορούσε να είναι: **«Ποιες οι διαφορές της γονιδιακής έκφρασης ανάμεσα στα κύτταρα των υγιών και των ασθενών οι οποίοι πάσχουν από σακχαρώδη διαβήτη;»**. Η διατύπωση του ερωτήματος παίζει πολύ σημαντικό ρόλο καθώς δίνει στην έρευνα ένα συγκεκριμένο στόχο, βοηθάει στην επιλογή κριτηρίων ελέγχου και καθοδηγεί την ανάλυση των δεδομένων και τη μοντελοποίησή τους.

Ένας άλλος πολύ σημαντικός παράγοντας που έχει καθοριστικό ρόλο στην εξέλιξη του πειράματος, είναι η επιλογή του βιολογικού υλικού το οποίο είναι διαθέσιμο. Αν για παράδειγμα το δείγμα που επιλέξουμε είναι mRNA, το οποίο είναι πιο ασταθές από το DNA. Επίσης, πρέπει να είναι γνωστή η ποσότητα του βιολογικού υλικού για να ενισχυθεί ή όχι η σήμανσή της.



Εικόνα 8: Βασικά βήματα για ένα πείραμα μικροσυστοιχιών (επιγραμματικά).

1.4.2 Επιλογή του κατάλληλου τύπου μικροσυστοιχίας:

Δεύτερο στάδιο αποτελεί η επιλογή του τύπου μικροσυστοιχίας. Η απόφαση για το καταλληλότερο τύπο μικροσυστοιχίας καθορίζεται από τη διαθεσιμότητα των μηχανημάτων καθώς και από τη συμβατότητα αυτών (των εργαλείων) με τον εκάστοτε τύπο (μικροσυστοιχίας). Πέρα όμως από τα παραπάνω, η επιλογή του τύπου μικροσυστοιχίας μπορεί να κατευθυνθεί άμεσα και από το βιολογικό ερώτημα το οποίο έχουμε θέσει.

Ανάλογα με το τρόπο κατασκευής τους, οι μικροσυστοιχίες διακρίνονται σε αυτές που κατασκευάστηκαν με *in situ* σύνθεση ανιχνευτών και σε μικροσυστοιχίες στις οποίες οι ανιχνευτές εκτυπώθηκαν στην επιφάνεια μετά την παρασκευή τους [Hongfang Liu, 2007]. Η κατηγορία των μικροσυστοιχιών στις οποίες οι ανιχνευτές εκτυπώθηκαν μετά τη παρασκευή τους χωρίζονται σε δύο άλλες κατηγορίες ανάλογα με μήκος των ανιχνευτών. Αυτές είναι οι μικροσυστοιχίες ολιγονουκλεοτιδίων και οι μικροσυστοιχίες cDNA.

1.4.2.1 *In situ* μικρών ολιγονουκλεοτιδικών αλληλουχιών:

Αυτού του τύπου οι μικροσυστοιχίες «gene-chip» παρασκευάζονται κυρίως από την εταιρεία της Affymetrix, και κατασκευάζονται με *in situ* σύνθεση μικρών ολιγονουκλεοτιδικών αλληλουχιών σε γυάλινο πλακίδιο με τη χρήση κατευθυνόμενο φωτός. Αυτό έχει σαν επακόλουθο την ακριβή παρασκευή ενός τακτοποιημένου, σε μεγάλο βαθμό, πίνακα DNA ολιγομερών στην επιφάνεια του πλακιδίου [Schna, 2003].

1.4.2.2 Μικροσυστοιχίες εκτύπωσης:

Όταν οι ανιχνευτές δε συντίθενται απευθείας πάνω στο πλακίδιο, αλλά εκτυπώνονται στην επιφάνεια μετά την ετοιμασία τους (spotted). Όπως ήδη έχει αναφερθεί οι μικροσυστοιχίες εκτύπωσης, διακρίνονται, ανάλογα με το είδος των ανιχνευτών σε ολιγονουκλεοτιδικές και cDNA μικροσυστοιχίες

Μικροσυστοιχίες ολιγονουκλεοτιδίων: με την αλματώδη αύξηση των γωιδιωματικών βάσεων δεδομένων καθώς και την ολοένα και αυξανόμενη διαθεσιμότητα «φθηνών» ολιγονουκλεοτιδίων, άρχισε να καλλιεργείται η τάση για χρήση επιμηκών ολιγονουκλεοτιδίων (50-70 nt) σαν ανιχνευτές για εφαρμογές γονιδιακής έκφρασης [Schena, 2003].

cDNA μικροσυστοιχίες: γνωστά ως και παράγωγα PCR για μικροσυστοιχίες, προκύπτουν με χρήση κοινών εκκινητών ή ειδικών-γονιδίου εκκινητών και τυπώνονται στην επιφάνεια της μικροσυστοιχίας. [Schena, 2003].

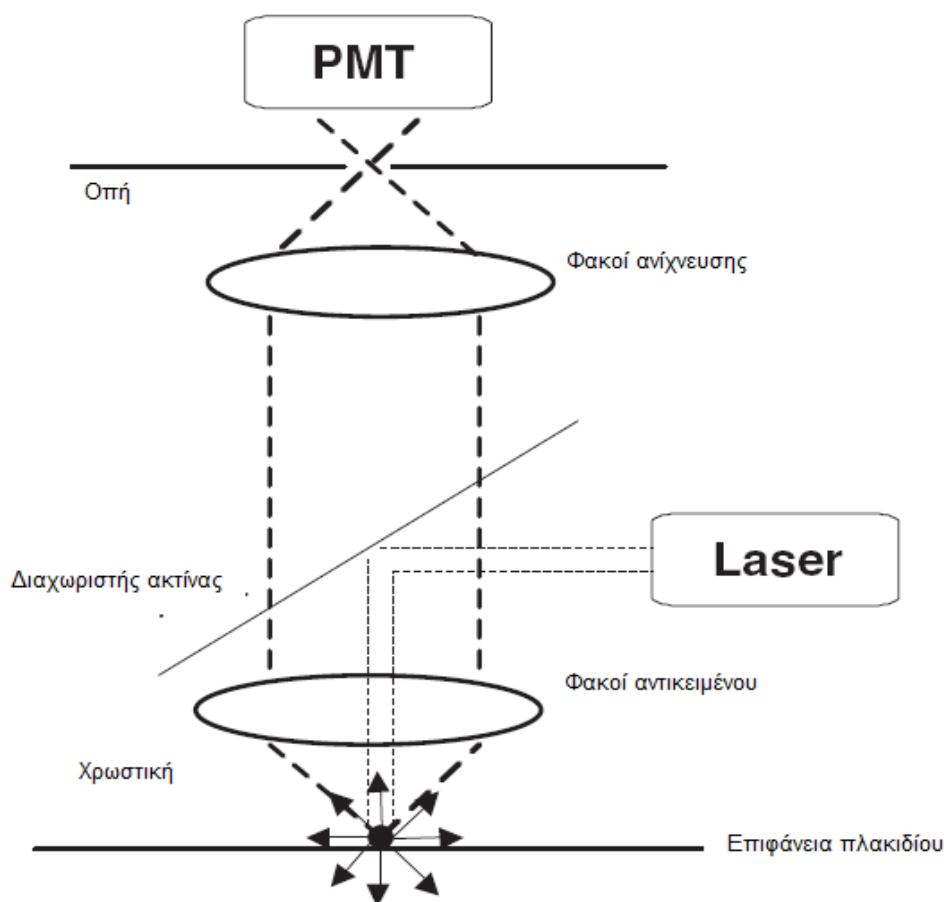
1.4.3 Παρασκευή δείγματος:

Σε ένα πείραμα μικροσυστοιχιών, το στάδιο που αποτελεί το κρίσιμότερο σημείο είναι αυτό της απομόνωσης του βιολογικού υλικού. Κατά την ανάλυση του γονιδιακού προφίλ έκφρασης γίνεται η απομόνωση RNA από κύτταρα ή ιστούς. Αφού γίνει η απομόνωση του mRNA, αυτό ιχνηθετείται (σημαίνεται).

Τόσο στις ολιγονουκλεοτιδικές όσο και στις cDNA μικροσυστοιχίες, χρησιμοποιούνται πάντα δύο δείγματα. Το ένα αναφέρεται ως δείγμα αναφοράς (πρόκειται για φυσιολογικό ιστό) και ιχνηθετείται με Cy3 και έχει χρώμα πράσινο και το άλλο αναφέρεται ως δείγμα ελέγχου (πρόκειται για ιστό που νοσεί) και ιχνηθετείται με Cy5 και έχει χρώμα κόκκινο. Και τα δύο αυτά δείγματα mRNA μεταγράφονται σε cDNA με τη βοήθεια της αντίστροφης μεταγραφάσης (ένζυμο) [Willard M. Freeman, 2000]. Κατόπιν, της διαδικασίας της ιχνηθέτησης (όπως έχει αναφερθεί παραπάνω) τα μόρια cDNA τοποθετούνται στις διατάξεις των μικροσυστοιχιών ώστε να υβριδοποιηθούν με τους ανιχνευτές. Για να εκτιμηθούν οι σχετικές ποσότητες για κάθε μετάγραφο των γονιδίων σε κάθε δείγμα, μετράται η ένταση του σήματος φθορισμού η οποία προκύπτει από τη διέγερση των ιχνηθετών στα αντίστοιχα μήκη κύματος.

1.4.4 Υβριδοποίηση:

Στη συνέχεια ακολουθεί το στάδιο της διαδικασίας της υβριδοποίησης του ιχνηθετημένου στόχου με τον αντίστοιχο ανιχνευτή. Τοποθετούνται οι στόχοι, οι οποίοι βρίσκονται σε κατάλληλο ρυθμιστικό διάλυμα, πάνω στη διάταξη της αλληλουχίας. Οι στόχοι με τους ανιχνευτές επωάζονται για συγκεκριμένο χρονικό διάστημα και με συγκεκριμένη θερμοκρασία. Η σύσταση της αλληλουχίας, το μήκος του ανιχνευτών και των στόχων, η θερμοκρασία υβριδοποίησης, η δευτεροταγής δομή, η συγκέντρωση αλάτων, το pH και αρκετοί άλλοι παράγοντες επηρεάζουν την αποτελεσματικότητα της υβριδοποίησης και τη συνοχή των διμερών που προκύπτουν. Κατόπιν, η μικροσυστοιχία πλένεται για να απομακρυνθούν τυχόν υβριδοποιήσεις οι οποίες δεν ήταν επιθυμητές (μη ειδικές) καθώς και το διάλυμα υβριδοποίησης. Έτσι αφού ληφθούν όλοι οι παράγοντες υβριδοποίησης υπόψη, το πλακίδιο είναι έτοιμο για να σαρωθεί από το ειδικό μηχάνημα και να παραχθεί η εικόνα.

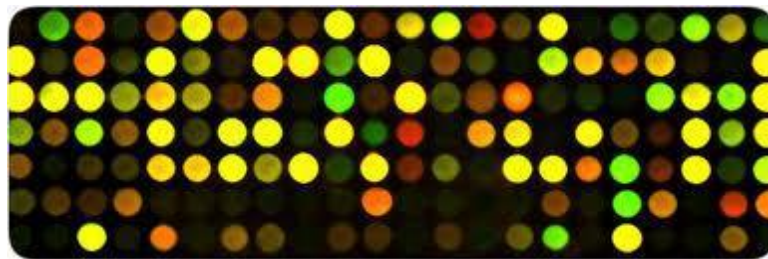


Εικόνα 9: Μια ακτίνα λέιζερ χρησιμοποιείται για να διεγείρει τις χρωστικές στην επιφάνεια της μικροσυτοιχίας. Στη συνέχεια, η ακτινοβολία που εκπέμπει η χρωστική συλλέγεται από τον φωτοπολλαπλασιαστή (PMT) και μετατρέπεται σε ψηφιακό σήμα.

1.4.5 Σάρωση:

Το τελευταίο κομμάτι που ολοκληρώνει τη διαδικασία πειράματος μικροσυτοιχιών είναι η σάρωση. Αφού τα δείγματα ιχνηθετούνται με διαφορετικό χρώμα, η εικόνα που θα δώσουν είναι διαφορετική και ανιχνεύεται σε διαφορετικό μήκος κύματος. Έτσι κατά τη διάρκεια της σάρωσης, ο σαρωτής παράγει δύο εικόνες, μία για κάθε χρώμα φθορισμού [Kerr, 2007]. Οι στόχοι που βρίσκονται σε πλεόνασμα από το κάθε δείγμα για κάθε γονίδιο, θα υπερισχύσουν έναντι των λιγότερων άλλων και θα καταλάβουν περισσότερους ανιχνευτές στο σημείο. Θα δούμε κόκκινο σημείο αν το συγκεκριμένο γονίδιο υπερεκφράζεται στα κύτταρα του πρώτου δείγματος, θα δούμε πράσινο σημείο αν ισχύει το αντίστοιχο για το δεύτερο δείγμα και, τέλος, θα δούμε κίτρινο σημείο αν η έκφραση είναι παρόμοια [Scheda, 2003].

Όταν, τέλος, έχουμε την εικόνα, προχωράμε σε ανάλυσή της, από την οποία προκύπτουν ποσοτικοποιημένα δεδομένα. Τα δεδομένα αυτά επεξεργάζονται στον ηλεκτρονικό υπολογιστή με πληθώρα αλγορίθμων, ώστε να απαλειφθούν τα σφάλματα και να δώσουν συμπεράσματα τα οποία ο άνθρωπος δε θα μπορούσε να εξάγει, λόγω του μεγάλου όγκου.



Εικόνα 10: Η συνδυασμένη εικόνα που παράγεται

Η συνδυασμένη εικόνα παρέχει έναν βολικό τρόπο να δει κανείς και να αναγνωρίσει τα γονίδια τα οποία βρίσκονται σε μεγαλύτερη έκφραση στο δείγμα ελέγχου σε σχέση με το δείγμα αναφοράς. Στην Εικόνα 10 μία κουκκίδα εμφανίζεται με αντίστοιχο χρώμα της ποσότητας του δείγματος ελέγχου και του δείγματος αναφοράς. Πιο συγκεκριμένα, οι ερμηνείες δίνονται ως εξής:

- ✓ Με κόκκινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος ελέγχου είναι μεγαλύτερο.
- ✓ Με πράσινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος αναφοράς είναι μεγαλύτερο.
- ✓ Με κίτρινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν οι ποσότητες του δείγματος ελέγχου και του δείγματος αναφοράς είναι ίσες.
- ✓ Με μαύρο χρώμα εμφανίζεται μία κουκκίδα αν κανένα δείγμα δεν έχει υβριδοποιηθεί.
- ✓ Οι υπόλοιπες αποχρώσεις εμφανίζονται για αντίστοιχες ποσότητες των δύο δειγμάτων [Kerr, 2007].

1.5 Ποσοτικοποίηση των δεδομένων:

Η συνδυασμένη εικόνα που προκύπτει από τη σάρωση εμπεριέχει όλα τα δεδομένα τα οποία χρειαζόμαστε για να εξάγουμε τα αποτελέσματα που επιθυμούμε. Κάθε κουκκίδα της εικόνας (pixel) αντιστοιχεί σε ένα γονίδιο. Σε κάθε κουκκίδα αντιστοιχούν πολλά ψηφία τα οποία ενσωματώνουν τις ποσοτικές πληροφορίες για το γονίδιο. Τα ψηφία αυτά είναι ανάλογα της έντασης του φθορισμού. Έτσι, τα αριθμητικά αυτά δεδομένα δίνουν τις πληροφορίες για την έκφραση των γονιδίων της μικροσυστοιχίας.

Το σχετικό επίπεδο έκφρασης για κάθε γονίδιο αντιστοιχεί στη ποσότητα του κόκκινου ή του πράσινου φωτός που εκπέμπεται μετά τη διέγερση. Για να συσχετίσουμε αυτές τις ποσότητες και να εξάγουμε το σχετικό επίπεδο έκφρασης κάθε γονιδίου χρησιμοποιούμε το λόγο έκφρασης. Είναι η πρώτη επεξεργασία που γίνεται στα δεδομένα για την ανίχνευση διαφορικής έκφρασης [Yang et al, 2002]. Αν έχουμε μια μικροσυστοιχία με $N_{\text{μικροσυστοιχία}}$ διακριτά στοιχεία, και συγκρίνουμε ένα δείγμα αναφοράς με ένα δείγμα προς εξέταση, τα οποία ονομάζουμε R και G (λόγω κόκκινης και πράσινης σήμανσης), τότε η αναλογία T για το i γονίδιο δίνεται από τον τύπο [Quackenbush, 2001]:

$$T_i = \frac{R_i}{G_i} \Leftrightarrow T'_i = \log_2(T_i)$$

Επειδή η αναλογία T δίνει μια μέτρηση των αλλαγών έκφρασης, έχει διαφορετική συμπεριφορά στα θετικά και αρνητικά ρυθμιζόμενα γονίδια. Για να εξαχθεί ένα συνεχές φάσμα τιμών και να υπάρχει παρόμοια συμπεριφορά στην επεξεργασία των θετικά και αρνητικά ρυθμιζόμενων γονιδίων, λογαριθμούμε με βάση το 2 το λόγο T_i .

1.6 Κανονικοποίηση:

Παρόλο που έχουμε εξάγει τα αριθμητικά μας δεδομένα και ενώ η διαδικασία φαίνεται να τελειώνει, ωστόσο πριν την ανάλυση αυτών των δεδομένων πρέπει να ληφθούν κάποιες παράγοντες υπόψη. Τα δεδομένα μας πρέπει να εξισορροπηθούν καθώς κατά τη διαδικασία του πειράματος γίνονται τυχαία και συστηματικά σφάλματα. Παραδείγματος χάριν, κατά τη διαδικασία της απεικόνισης μπορούν να προκύψουν ανισορροπίες που οφείλονται από διαφορές στην αποτελεσματικότητα της σήμανσης, της υβριδοποίησης, των εκπλύσεων καθώς και από διαφοροποιήσεις στην ένταση του λείζερ και την ευαισθησία του σαρωτή (εισαγωγή θορύβου στα δεδομένα). Επίσης λόγω διαφορετικών πειραματικών συνθηκών, η επανάληψη ενός πειράματος μπορεί να έχει διαφορετικά αποτελέσματα διακύμανσης στα σχετικά επίπεδα έκφρασης των γονιδίων, όπως και η χρήση διαφορετικών τύπων πλατφορμών. Έτσι προς αποφυγή των παραπάνω σφαλμάτων, κανονικοποιούμε τα δεδομένα μας.

Με τη κανονικοποίηση, τα δεδομένα εξισορροπούνται με τέτοιο τρόπο ώστε να μην παραποιούνται. Επιτυγχάνεται η ελαχιστοποίηση των συστηματικών λαθών που εντοπίζονται στα επίπεδα έκφρασης των γονιδίων και επιτρέπεται η σύγκρισή τους με άλλα δεδομένα που προέρχονται από διαφορετικού τύπου chip [Quackenbush, 2001]. Υπάρχουν 2 τρόποι ελαχιστοποίησης των σφαλμάτων στα επίπεδα έκφρασης των γονιδίων: α) Κανονικοποίηση ολικής έντασης και β) Lowess (locally weighted linear regression) κανονικοποίηση.

1.6.1 Κανονικοποίηση ολικής έντασης:

Για την εφαρμογή της βασιζόμαστε σε 2 υποθέσεις-παραδοχές. Η πρώτη υπόθεση: ότι η ποσότητα του mRNA και στα δύο δείγματα είναι ίδια άρα και κατ' επέκταση ίσος αριθμός μορίων για κάθε δείγμα και η δεύτερη υπόθεση: ότι υβριδοποιείται ίδιος αριθμός μορίων RNA και στα δύο δείγματα, άρα κατά συνέπεια η ένταση φθορισμού και των δύο δειγμάτων θα πρέπει να είναι ίσες μεταξύ τους [Quackenbush, 2001].

Αθροίζοντας τις εντάσεις και από τα δύο δείγματα, υπολογίζεται ένας παράγοντας κανονικοποίησης με τον εξής τρόπο [Quackenbush, 2001].

$$N_{ολικό} = \frac{\sum_{i=1}^{N_{\text{συστοιχία}}} R_i}{\sum_{i=1}^{N_{\text{συστοιχία}}} G_i}$$

Όπου:

R_i : μετρούμενη ένταση για το κόκκινο για το στοιχείο i της μικροσυστοιχίας

G_i : μετρούμενη ένταση για το πράσινο για το στοιχείο i της μικροσυστοιχίας

$N_{\text{συστοιχία}}$: ολικός αριθμός των στοιχείων της μικροσυστοιχίας

Οι εντάσεις ισοσταθμίζονται ως εξής:

$$G'_k = G_k * N_{\text{ολικό}}$$

$$R'_k = R_k$$

Και η κανονικοποιημένη πλέον αναλογία για κάθε στοιχείο δίνεται από το τύπο:

$$T_i = \frac{R_i}{G_i} = \frac{1}{N_{\text{ολικό}}} * \frac{R_i}{G_i}$$

Και προσαρμόζεται με τέτοιο τρόπο ώστε η μέση αναλογία να είναι ίση με τη μονάδα.

1.6.2 Lowess (γραμμική οπισθοδρόμηση τοπικών βαρών) κανονικοποίηση:

Η μέθοδος αυτή κανονικοποίησης μπορεί να εξαλείψει συστηματικές παρεκκλίσεις που ενδέχεται να έχουν οι λογαριθμικές αναλογίες και που οφείλονται στην ένταση του σήματος [Hoheisel JD, 2006]. Δημιουργώντας ένα διάγραμμα με άξονες το $\log_2(\frac{R_i}{G_i})$ και το $\log_{10}(R_i * G_i)$, μπορούμε να οπτικοποιήσουμε τις παρεκκλίσεις που οφείλονται στην ένταση του σήματος. Το διάγραμμα αυτό ονομάζεται R-I(ratio-intensity). Η Lowess μπορεί να ανιχνεύσει τις παρεκκλίσεις στο διάγραμμα και κάνοντας τοπική γραμμική οπισθοδρόμηση σε συνάρτηση των εντάσεων, να τις διορθώσει. Αυτό πραγματοποιείται κάνοντας χρήση μίας συνάρτησης βαρών η οποία κάνει λιγότερο σημαντική τη συμβολή των δεδομένων σημείων της μικροσυστοιχίας που απέχουν πολύ από κάθε σημείο, δηλαδή εξαλείφει την υπολογισμένα πιο ταιριαστή μέση αναλογία από τη πειραματικά παρατηρούμενη αναλογία για κάθε σημείο δεδομένου [Yang et al, 2002].

Πιο συγκεκριμένα, αν θέσουμε

$$x_i = \log_{10}(R_i * G_i) \text{ και } y_i = \log_2(\frac{R_i}{G_i})$$

η μέθοδος θα υπολογίσει πρώτα τη $y(x_k)$, την εξάρτηση δηλαδή της $\log_2(\text{αναλογίας})$ από τις $\log_{10}(\text{εντάσεις})$, και έπειτα με τη συνάρτηση θα διορθώσει σημείο προς σημείο τις τιμές της $\log_2(\text{αναλογίας})$ έτσι ώστε:

$$\log_2(T'_k) = \log_2(T_k) - y(x_k) = \log_2(T_k) - \log_2(2^{y(x_k)})$$

2 Κεφάλαιο:

2.1 Ανάλυση μικροσυστοιχιών-Στατιστική ανάλυση:

Καθώς τα πειράματα μικροσυστοιχιών υπολογίζουν τη ταυτόχρονη έκφραση χιλιάδων γονιδίων υπό συγκεκριμένες συνθήκες, τα δεδομένα απαιτούν μια στατιστική επιλογή έτσι ώστε να εντοπιστούν αυτά που εκφράζονται διαφορετικά και είναι στατιστικώς σημαντικά. Για τον εντοπισμό αυτών των γονιδίων υπάρχουν διάφοροι μέθοδοι στατιστικής ανάλυσης. Παραμετρικά τεστ, γνωστά ως t-tests, συγκρίνουν τις ομάδες (συνήθως δύο), ταυτόχρονα. Άλλες στατιστικές μέθοδοι που εφαρμόζονται είναι: η αλλαγή διπλώματος (fold change), η μέθοδος permute και η μέθοδος bootstrap.

2.1.1 T-test:

Το t-test αξιολογεί αν τα μέσα των δύο ομάδων έχουν στατιστικά σημαντικές διαφορές. Αυτή η ανάλυση χρησιμοποιείται για να συγκρίνει τα μέσα δύο ομάδων, και ειδικά για την ανάλυση των δύο ομάδων μόνο με τυχαία πειραματικό σχεδιασμό. Ελέγχει κατά πόσο η διαφορά μέσων τιμών ανάμεσα στους δύο πληθυσμούς μπορεί να προέκυψε από τύχη κατά την επιλογή του δείγματος.

Στις μελέτες μικροσυστοιχίας συνήθως υπάρχει μικρό μέγεθος δείγματος και μη κανονική κατανομή των τιμών έκφρασης. Έτσι, ο στατιστικός έλεγχος t-test συνοδεύεται από τη μέθοδο permute ή τη μέθοδο bootstrap, ώστε να παραχθούν εκ νέου τυπικό σφάλμα και πιο αξιόπιστα διαστήματα εμπιστοσύνης. Οι μέθοδοι Bootstrap και μετάθεσης (permutation) είναι δημοφιλείς μέθοδοι επαναδειγματοληψίας άμεσα διαθέσιμες σε μεγάλα στατιστικά πακέτα όπως το Stata και το R. Υπάρχουν διάφορες υλοποιήσεις της Bootstrap που διατίθεται τόσο στο Stata (bootstrap εντολή) όσο και στο R (εντολή boot).

2.1.2 Μέθοδος Permutation:

Η μέθοδος **permutation** είναι ένας έλεγχος στατιστικής σημαντικότητας, όπου υπολογίζει τη δειγματική κατανομή κάθε στατιστικού ελέγχου. Κάνοντας αναδιατάξεις των ετικετών (labels), υπολογίζει όλες τις πιθανές τιμές του στατιστικού ελέγχου στα παρατηρούμενα δεδομένα. Υπό την μηδενική υπόθεση της μη σύνδεσης, κάθε μετάθεση των ετικετών για τα δείγματα (case_control) θεωρείται πιθανή και αντιπροσωπεύει ένα τυχαίο περιστατικό των δεδομένων. 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, ωστόσο απαιτεί πολλές επαναλήψεις και δεν δουλεύει καλά για ζευγαρωτά δεδομένα.

2.1.3 Μέθοδος Bootstrap:

Η **Bootstrap** είναι μέθοδος που βασίζεται στην επαναδειγματοληψία και χρησιμοποιείται για τον υπολογισμό της δειγματικής κατανομής στατιστικών χωρίς τη χρήση της κανονικής θεωρίας (π.χ. t-test). Ως μέθοδος επαναδειγματοληψίας δίνει πιο ακριβείς απαντήσεις για το τυπικό σφάλμα και για τα διαστήματα εμπιστοσύνης. Έτσι είναι ιδιαίτερα αποτελεσματική στην διαδικασία της μετά-ανάλυσης των μικροσυστοιχιών, όταν η κατανομή είναι άγνωστη και το μέγεθος του δείγματος είναι μικρό. Η μέθοδος αποτελείται από τρία στάδια. Πρώτα γίνεται η αναδειγματοληψία, έπειτα υπολογίζεται η κατανομή της Bootstrap και τέλος η κατανομή που προκύπτει χρησιμοποιείται για την εξαγωγή των συμπερασμάτων.

Για να γίνει αντιληπτός ο τρόπος με τον οποίο δουλεύει η μέθοδος, έστω ότι υπάρχει ένα σύνολο δεδομένων N . Μετά, από τυχαία επαναδειγματοληψία και επανατοποθέτηση από το αρχικό πραγματικό δείγμα, εκτελείται ο αλγόριθμος «monte carlo» ο οποίος δημιουργεί χιλιάδες νέα δείγματα ίδιου μεγέθους με το αρχικό σύνολο δεδομένων. Εναλλακτικά πρέπει να υπολογιστούν 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 παράμετρος που επιθυμείται (π.χ. ο μέσος όρος του πληθυσμού) και δημιουργείται μια εκτίμηση της κατανομής της στατιστικής παραμέτρου, η Bootstrap κατανομή. Χρησιμοποιώντας την Bootstrap κατανομή λαμβάνονται πληροφορίες για το σχήμα, το κέντρο, το τυπικό σφάλμα και την τυπική απόκλιση της δειγματικής κατανομής.

Υπάρχουν διάφορες προσεγγίσεις της μεθόδου bootstrap. Οι πιο γνωστές είναι:

- **Bayesian Bootstrap:** οι παρατηρήσεις που επιλέγονται για δειγματοληψία και επανάθεση δεν έχουν όλες την ίδια πιθανότητα $\frac{1}{n}$ να επιλεγούν, αλλά διαφορετική.
- **Παραμετρική μέθοδος:** χρησιμοποιεί παραμετρικό μοντέλο κατά το στάδιο της δειγματοληψίας.
- **Ομαλοποιημένη μέθοδος:** υπολογίζεται μια «ομαλή εκδοχή» της εκτιμήτριας της δειγματικής κατανομής, με τη χρήση μεθόδων εξομάλυνσης πυρήνων.
- **M_out_of_N bootstrap:** επιλέγονται μικρότερα σε μέγεθος δείγματα bootstrap από το αρχικό ($m < n$), ώστε να υπάρχει μικρότερη εξάρτηση και η διασπορά της εκτιμήτριας να προσομοιάζεται καλύτερα.
- **Double Bootstrap:** δημιουργεί B^2 δείγματα, όπου B είναι τα bootstrap δείγματα που λαμβάνονται από το αρχικό δείγμα αλλά και αυτά που χρησιμοποιούνται για προσαρμογή στην εκτιμήτρια. Απαιτεί μεγάλη υπολογιστική ισχύ.

Από την εξαγωγή της bootstrap κατανομής, υπολογίζονται τα διαστήματα εμπιστοσύνης. Ο υπολογισμός τους γίνεται με διάφορους μεθόδους εύρεσης, ανάλογα με το αν η κατανομή είναι μεροληπτική και/ή αν έχει έλλειψη συμμετρίας. Η πιο γνωστή μέθοδο είναι:

- **Μέθοδος percentile:** χρησιμοποιεί τα ποσοστιαία σημεία της κατανομής bootstrap μιας στατιστικής συνάρτησης. Αν έχω μία παράμετρο θ , η εκτιμήτρια συνάρτησή της θα είναι $\hat{\theta}$, και η τιμή της στατιστικής συνάρτησης για το b bootstrap δείγμα θα είναι $\hat{\theta}(b)$. Άρα για τις πρώτες α πειρες επαναλήψεις, το πρώτο percentile διάστημα εμπιστοσύνης θα είναι

$[\hat{\theta}(\frac{\alpha}{2}), \hat{\theta}(1 - \frac{\alpha}{2})]$, ενώ για τις B επαναλήψεις: $[\hat{\theta}(\frac{\alpha}{2} * B), \hat{\theta}(1 - \frac{\alpha}{2} * B)]$ [Efron B., 1993].

- **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:

$$\alpha_1 = \Phi\left(\widehat{Z}_0 + \frac{\widehat{Z}_0 + Z_{\frac{\alpha}{2}}}{1 - \hat{\alpha}(\widehat{Z}_0 + Z_{\frac{\alpha}{2}})}\right) \text{ και}$$

$$\alpha_2 = \Phi\left(\widehat{Z}_0 + \frac{\widehat{Z}_0 + Z_{1 - \frac{\alpha}{2}}}{1 - \hat{\alpha}(\widehat{Z}_0 + Z_{1 - \frac{\alpha}{2}})}\right)$$

με Z_{α} : 100^α ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής

Φ: τυποποιημένη κανονική αθροιστική κατανομή [Thomas J. DiCiccio & Efron, 1996].

- **Μέθοδος με κανονική προσέγγιση:** Αν έχω μία παράμετρο θ με κατανομή F , και η εκτιμήτριά της $\hat{\theta} = T^*$ με κατανομή F_T , τότε για να ισχύει

$$P(c_{1 - \frac{\alpha}{2}} < \hat{\theta} - \theta < c_{\frac{\alpha}{2}}) = 1 - \alpha$$

το διάστημα εμπιστοσύνης θα είναι: $[\hat{\theta} - c_{\frac{\alpha}{2}}, \hat{\theta} - c_{1 - \frac{\alpha}{2}}]$. Αυτή η μέθοδος πλεονεκτεί καθώς παρέχει ένα απλό τρόπο υπολογισμού των διαστημάτων εμπιστοσύνης, ειδικά αν μελετάται η μέση τιμή.

2.1.4 Μέθοδοι διόρθωσης του p-value:

Με τον έλεγχο του t-test, προκύπτει ένα συγκεκριμένο p-value για κάθε γονίδιο, το οποίο μας υποδεικνύει αν είναι στατιστικά σημαντικό ή όχι. Το σύνηθες επίπεδο σημαντικότητας για το p-value είναι $p < 0.05$ και $p < 0.01$. Ωστόσο, παρουσιάζονται προβλήματα κατά την επιλογή του p-value, καθώς ανάλογα με το όριο υπάρχει και η ανάλογη εμφάνιση λανθασμένων θετικά σημαντικών γονιδίων (false positive). Ειδικότερα στα πειράματα μικροσυστοιχιών, απαιτούνται μέθοδοι διόρθωσης του p-value καθώς τα επίπεδα σημαντικότητας (5% και 1%) περιέχουν λανθασμένα στατιστικώς σημαντικά γονίδια, καθώς ο αριθμός των γονιδίων είναι πολύ μεγάλος. Με τις μεθόδους διόρθωσης, δίνεται η δυνατότητα επιλογής του επιπέδου σημαντικότητας, έτσι κάθε p-value κάθε γονιδίου αναπροσαρμόζεται ώστε το ποσοστό των λανθασμένων θετικά σημαντικών γονιδίων να βρίσκεται κάτω από το προκαθορισμένο όριο.

2.1.4.1 Διόρθωση Bonferroni:

Όταν διεξάγονται ταυτόχρονα πολλές ανεξάρτητες ή εξαρτημένες στατιστικές δοκιμές, η μέθοδος διόρθωσης του Bonferroni είναι αποτελεσματική καθώς επαναπροσδιορίζει το επίπεδο σημαντικότητας. Μετατρέπει στα τεστ ελέγχου το επίπεδο σημαντικότητας α σε $\alpha_{new} = \frac{\alpha}{n}$, όπου n : ο αριθμός δειγμάτων. Έτσι αν το $p\text{-value} < \alpha_{new}$, τότε είναι στατιστικά σημαντικό [Eye, 2003]. Οι διορθωμένες τιμές δίνονται από το τύπο: $p_{cor(i)} = n * p_{(i)}$ (όταν το γινόμενο υπερβαίνει το 1, η τιμή γίνεται 1).

2.1.4.2 Benjamini and Hochberg false discovery rate (FDR):

Η μέθοδος διόρθωσης FDR, παρότι είναι λιγότερο συντηρητική διαδικασία από τη διόρθωση Bonferroni, χρησιμοποιείται συχνά στις μικροσυστοιχίες καθώς δίνει μία καλή ισορροπία των στατιστικά σημαντικών γονιδίων και αυτών που είναι λανθασμένα στατιστικώς σημαντικά. Αρχικά, οι τιμές των $p\text{-value}$ ταξινομούνται κατά αύξουσα σειρά. Έπειτα υπολογίζεται το γινόμενο $\frac{N * p\text{-value}}{k}$ και της προσωρινής τιμής, όπου N : ο αριθμός των πολλαπλών ερωτημάτων και k : η σειρά. Σαν προσωρινή τιμή ορίζεται η μονάδα(1). Αρχίζοντας από το μεγαλύτερο $p\text{-value}$, κάθε διορθωμένη τιμή είναι η μικρότερη μεταξύ του γινομένου $\frac{N * p\text{-value}}{k}$ και της προσωρινής τιμής (για την τήρηση της μονοτονίας, μετά το τέλος κάθε διόρθωσης, προσωρινή τιμή γίνεται η διορθωμένη). Η διορθωμένη τιμή για τα $p\text{-value}$ (εκτός του μεγαλύτερου) δίνεται από το τύπο: $p_{cor(i)} = \frac{n}{n-i} * p_{(i)}$ [Benjamini and Hochberg, 1995].

2.1.4.3 Διόρθωση Bonferroni step down(Holm):

Είναι παρόμοια μέθοδος με του Bonferroni, αλλά πιο αυστηρή όσον αφορά το νέο $p\text{-value}$. Οι τιμές των $p\text{-value}$ ταξινομούνται κατά αύξουσα σειρά [Holm, 1979]. Το πρώτο $p\text{-value}$ πολλαπλασιάζεται με το συνολικό αριθμό γονιδίων (έστω n) της μικροσυστοιχίας και όλα τα υπόλοιπα $p\text{-value}$ υπόκεινται διόρθωση σύμφωνα με το τύπο: $p_{cor(i)} = (n - i) * p_{(i)}$.

2.1.4.4 Διόρθωση Holland:

Τα βήματα είναι ακριβώς τα ίδια με αυτά της διόρθωσης Holm, με διαφορά ότι το πρώτο $p\text{-value}$ πολλαπλασιάζεται με $(n-1)$ συνολικό αριθμό γονιδίων και ο τύπος της διόρθωσης των υπολοίπων $p\text{-value}$ δίνεται από το τύπο: $p_{cor(i)} = (n - i + 1) * p_{(i)}$ [Holland and Copenhaver, 1987].

2.1.4.5 Διόρθωση Sidak:

Είναι μία απλή μέθοδος και προτιμάται από τη διόρθωση Bonferroni. Χρησιμοποιείται στους πολλαπλούς ελέγχους και ο τύπος διόρθωσης των p-value δίνεται :

$$p_{cor(i)} = 1 - (1 - p_{(i)})^{\frac{1}{n}}, [\text{Šidák, 1967}].$$

2.2 Ανάλυση μικροσυστοιχιών-Ομαδοποίηση(Clustering):

Η ομαδοποίηση αποτελεί το δεύτερο στάδιο της ανάλυσης των μικροσυστοιχιών. Ανάλογα με τα επίπεδα έκφρασης, τα γονίδια ομαδοποιούνται. Έπειτα γίνεται αναπαράσταση των ομάδων με σκοπό την εύρεση πιθανών σχέσεων μεταξύ των γονιδίων.

Οι αλγόριθμοι ομαδοποίησης είναι ευρέως διαδεδομένοι και εφαρμόσιμοι σε πειράματα μικροσυστοιχιών. Έχουν προταθεί διάφορες τεχνικές ομαδοποίησης για εξαγωγή προτύπων γονιδιακής έκφρασης [Sturn et al., 2002].

Χωρίζονται σε **ιεραρχικούς** και **μη ιεραρχικούς** αλγορίθμους. Στη περίπτωση των **ιεραρχικών αλγορίθμων** η ταξινόμηση που προκύπτει έχει έναν αυξανόμενο αριθμό εμφωλευμένων κλάσεων και το αποτέλεσμα μοιάζει με φυλογενετική ταξινόμηση. Στους **μη ιεραρχικούς**, όπως ο k _means, απλά διαχωρίζουν τα αντικείμενα σε διαφορετικές ομάδες, χωρίς να προσπαθούν να βρουν τη σχέση μεταξύ των ξεχωριστών αντικειμένων [Leung & Cavalieri, 2003].

Επίσης, μπορούν να χωριστούν σε **διαχωριστικούς** (divisive) και **συνενωτικούς** (agglomerative). Οι **διαχωριστικοί** ξεκινούν με όλα τα δεδομένα σε μια ομάδα, τα οποία σταδιακά χωρίζονται σε όλο και μικρότερες. Την αντίθετη λειτουργία κάνουν **οι συνενωτικοί αλγόριθμοι** οι οποίοι ξεκινούν από ομάδες του ενός μόνο στοιχείου και σταδιακά δημιουργούν ομάδες με περισσότερα μέλη.

Ένας ακόμη διαχωρισμός των αλγορίθμων ομαδοποίησης είναι σε **επιβλεπόμενους** (supervised) και **μη επιβλεπόμενους** (unsupervised) αλγορίθμους. Στους **επιβλεπόμενους**, ο χρήστης μπορεί να καθορίσει τον αριθμό των ομάδων που επιθυμεί ενώ στους **μη επιβλεπόμενους** ο αλγόριθμος, ανάλογα με τα δεδομένα, παράγει αυτόματα τον αριθμό των ομάδων [Quackenbush, 2001].

Όλοι οι αλγόριθμοι ομαδοποίησης λαμβάνουν σαν είσοδο, την απόσταση δύο γονιδίων. Ο τρόπος υπολογισμού της απόστασης παίζει σημαντικό ρόλο καθώς δίνει διαφορετικά αποτελέσματα ανά περίπτωση.

Υπάρχουν τρία βασικά είδη υπολογισμού αποστάσεων μεταξύ γονιδίων. Η **ευκλείδεια απόσταση** η οποία υπολογίζεται βάση του τύπου: $d_{AB} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$, όπου x_i & y_i : οι υπολογισμένες τιμές της έκφρασης των γονιδίων A&B αντίστοιχα και το συνολικό άθροισμα: τα n πειράματα που μελετώνται.

Η **απόσταση Manhattan**, υπολογίζεται από το τύπο: $d_{AB} = \sum_{i=1}^n |x_i - y_i|$, όπου x_i & y_i : οι υπολογισμένες τιμές της έκφρασης των γονιδίων A&B αντίστοιχα και το συνολικό άθροισμα: τα n πειράματα που μελετώνται.

Τέλος υπάρχει ο **συντελεστής συσχέτισης του Pearson**, που δίνεται από το τύπο:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}}, \text{ όπου}$$

x_i & y_i : οι υπολογισμένες τιμές της έκφρασης των γονιδίων, το συνολικό άθροισμα: τα n πειράματα που μελετώνται και \bar{X} & \bar{Y} : η μέση τιμή έκφρασης όλων των τιμών έκφρασης των γονιδίων σε όλα τα πειράματα.

2.3 Ανάλυση μικροσυστοιχιών-Πρόγνωση:

Η πρόγνωση είναι το τελευταίο κομμάτι που συμπληρώνει την ανάλυση μικροσυστοιχιών. Μας ενδιαφέρει κυρίως η σωστή πρόγνωση (ταξινόμηση) των ασθενών. Σε περιπτώσεις πρόβλεψης της ασθένειας, έχει σημασία ως διαγνωστική δοκιμασία. Υπάρχει ποικιλία μεθόδων που κάνουν ταξινόμηση. Οι πιο συνηθισμένες είναι τα νευρωνικά δίκτυα και η support vector machine(SVM). Ωστόσο, πολλές φορές απαιτείται κάποια μέθοδος επιλογής των πιο σημαντικών γονιδίων.

2.3.1 Ο αλγόριθμος ιεραρχικής ταξινόμησης:

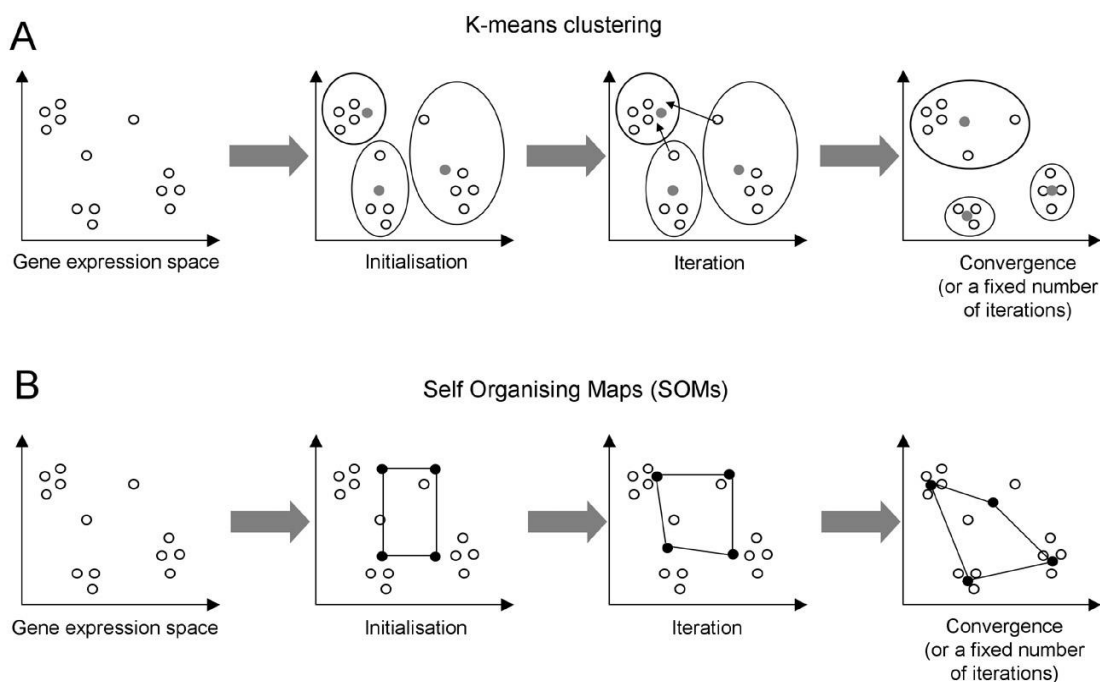
Ο αλγόριθμος αυτός χρησιμοποιείται ευρέως στην ανάλυση δεδομένων γονιδιακής έκφρασης. Θεωρείται απλός και δίνει τη δυνατότητα της αναπαράστασης των αποτελεσμάτων γραφικά. Η φιλοσοφία του βασίζεται στο ότι κάθε αντικείμενο αποτελεί και μία ξεχωριστή συστάδα. Με τη χρήση της ευκλείδειας απόστασης, υπολογίζονται οι αποστάσεις ανά δύο αντικείμενων. Κατόπιν, αφού ενωθούν τα αντικείμενα τα οποία είναι κοντινότερα το ένα στο άλλο, σχηματίζουν μία νέες συστάδες. Για τις νέες συστάδες που δημιουργούνται, υπολογίζονται και πάλι οι αποστάσεις αυτών από τις άλλες. Η διαδικασία επαναλαμβάνεται τόσες φορές όσες χρειάζεται για να καταλήξει τελικά ο αλγόριθμος σε μία συστάδα όπου θα περιέχονται όλα τα αντικείμενα.

Υπάρχουν κατηγορίες αλγορίθμων ταξινόμησης, οι οποίοι διαφοροποιούνται στο τρόπο υπολογισμού των αποστάσεων μεταξύ των ομάδων [Quackenbush, 2001]. Τέτοιοι είναι:

- **Ιεραρχική ταξινόμηση απλής σύνδεσης (single linkage clustering):** η απόσταση δύο ομάδων είναι η ελάχιστη απόσταση ενός μέλους της μίας ομάδας από ένα μέλος της άλλης ομάδας. Η μέθοδος αυτή είναι γνωστή και σαν κοντινότερου γείτονα ή μέθοδος ελαχίστου.
- **Ιεραρχική ταξινόμηση ολικής σύνδεσης (complete linkage clustering):** η απόσταση δύο ομάδων, αυτή τη φορά είναι η μέγιστη απόσταση ενός μέλους της μίας ομάδας από ένα μέλος της άλλης ομάδας. Η μέθοδος είναι γνωστή σαν του απόμακρου γείτονα ή μέθοδος μεγίστου.
- **Ιεραρχική ταξινόμηση μέσης σύνδεσης (average linkage clustering):** γίνεται χρήση των μέσων όρων των αντικειμένων που ανήκουν σε μία συστάδα. Πιο συγκεκριμένα, η απόσταση δύο συστάδων είναι ίση με τη μέση τιμή της απόστασης κάθε αντικειμένου της μίας με όλα τα αντικείμενα της άλλης συστάδας.

2.3.2 Οι διαχωριστικοί αλγόριθμοι ομαδοποίησης:

- **Μέθοδος k-means:** αρχικά ορίζεται ο αριθμός των επιθυμητών ομάδων, τα δεδομένα χωρίζονται τυχαία μέσα σε αυτές και ορίζονται τα κεντροειδή των ομάδων. Κάθε σημείο μεταβιβάζεται στο κεντροειδές της κοντινότερης ομάδας. Επανα-υπολογίζονται τα κεντροειδή των νέων ομάδων. Ο αλγόριθμος συνεχίζει μέχρι να συγκλίνει σε κάποιο κριτήριο
- **SOMs-self organizing maps:** βασίζεται στα νευρωνικά δίκτυα και προτάθηκε από το Kohonen το 1980. Η φιλοσοφία των νευρωνικών δικτύων είναι ότι εκπαιδεύονται από τα ίδια τα δεδομένα τους και από τις εξόδους που αντιστοιχούν σε αυτά με τέτοιο τρόπο ώστε να αντιμετωπίζουν τα εκάστοτε προβλήματα. Ο αλγόριθμος αυτός αποτελείται από ένα επίπεδο νευρώνων εισόδου και ένα επίπεδο ανταγωνιστικών νευρώνων στους οποίους αντιστοιχούν διανύσματα βαρών. Όταν εφαρμοστεί είσοδος στο σύστημα, οι νευρώνες ανταγωνίζονται μεταξύ τους. Νικητής θεωρείται εκείνος ο νευρώνας όπου το διάνυσμα βαρών του ομοιάζει περισσότερο με την είσοδο [Dalton L, 2009].
- **PCA-principle component analysis:** πραγματοποιεί έρευνα σε πολλές διαστάσεις και επιτρέπει στα σύνθετα δεδομένα να απεικονίζονται σε μικρότερο χώρο διατηρώντας την απόκλιση τους.



Εικόνα 11: Α) Σχηματική αναπαράσταση της φιλοσοφίας της ομαδοποίησης k-means Β) Σχηματική αναπαράσταση της αρχής στην οποία βασίζεται η ομαδοποίηση των SOMs, πηγή: Madam Badu et al.

2.3.3 Μέθοδοι επιβλεπόμενης μάθησης:

Support vector machine (SVM): εκπαιδεύονται από έναν αρχικό αριθμό δεδομένων και έπειτα μπορούν να αναγνωρίσουν και να ξεχωρίσουν τα υπόλοιπα με βάση την έκφραση τους. Το πλεονέκτημά τους είναι ότι χρησιμοποιούν βιολογικές πληροφορίες για να καθορίσουν τα στοιχεία έκφρασης που χρειάζονται για να χαρακτηρίσουν τις ομάδες και μετά, είναι ικανοί να κατηγοριοποιήσουν οποιοδήποτε άλλο αντικείμενο.

2.3.4 Μέθοδοι μη επιβλεπόμενης μάθησης:

Ο αλγόριθμος **MCL(Markov clustering)**: παίρνει σαν είσοδο μία λίστα από ζευγάρια τιμών με το αντίστοιχο βάρος τους ενώ ο αριθμός των ομάδων ορίζεται αυτόματα. Έχει χρησιμοποιηθεί για την ομαδοποίηση πρωτεϊνικών ακολουθιών, τη πρόγνωση πρωτεϊνικών οικογενειών και συμπλόκων καθώς, πλέον είναι εφαρμόσιμος και στη τεχνολογία μικροσυστοιχιών για την ομαδοποίηση των γονιδίων [B. Samuel Lattimorea, 2005].

2.4 Βάσεις δεδομένων γονιδιακής έκφρασης:

Με τη δημιουργία νέων τεχνολογιών όπως των Next Generation Sequencing (NGS) καθώς και τη μεγάλη γκάμα διάθεσης οικονομικότερων τσιπ μικροσυστοιχιών, η ανάλυση δεδομένων γονιδιακής έκφρασης τα τελευταία χρόνια έχει επεκταθεί πολύ. Ως αποτέλεσμα αυτού, προκύπτει η ανάγκη για διαχείριση και επεξεργασία μεγάλου όγκου δεδομένων. Έτσι, προς ικανοποίηση της ανάγκης αυτής, δημιουργήθηκαν οι βάσεις δεδομένων γονιδιακής έκφρασης, οι οποίες δίνουν τη δυνατότητα καταχώρησης των δεδομένων ενός πειράματος ενώ μερικές από αυτές παρέχουν επιπλέον εργαλεία ανάλυσης. Επίσης, παρέχουν και πληροφορίες σχετικά με τη πλατφόρμα που χρησιμοποιήθηκε, τα γονίδια που μελετήθηκαν, τα δείγματα που έλαβαν μέρος στο πείραμα και το είδος των δεδομένων.

Όπως είναι αναμενόμενο, η μεγάλη πολυπλοκότητα και ο τεράστιος όγκος αυτών των δεδομένων για να μπορούν να γίνουν εύκολα επεξεργάσιμα αλλά και για να μπορούν να διατεθούν σε κάποια δημόσια βάση δεδομένων, θα πρέπει να ακολουθούν κάποιο πρωτόκολλο. Για το σκοπό αυτό έχει καθιερωθεί ένα συγκεκριμένο πρωτόκολλο βάσει του οποίου καταχωρείται η ελάχιστη πληροφορία που περιγράφει ένα πείραμα μικροσυστοιχιών. Το πρωτόκολλο αυτό ονομάζεται **MIAME: Minimum Information About Microarray Experiment**, και γίνεται προσπάθεια «επιβολής» του στους συγγραφείς που επρόκειτο να δημοσιεύσουν κάποια σχετική εργασία. Για να γίνει η εργασία αποδεκτή πρέπει πρώτα να έχουν κατατεθεί τα δεδομένα του πειράματος σε κάποια δημόσια βάση δεδομένων. Οι πιο διαδεδομένες βάσεις είναι:

- **GEO(Gene Expression Omnibus)**: αποτελεί βάση δεδομένων του NCBI: National Center for Biotechnology Information. Περιέχει δεδομένα γονιδιακής έκφρασης από μικροσυστοιχίες αλλά και από Next Generation Sequencing (NGS). Η επίσημη ιστοσελίδα της είναι: <http://www.ncbi.nlm.nih.gov/geo/> [Barrett T& Edgar, 2006].
- **Array Express**: αποτελεί βάση δεδομένων του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EBI). Δεν διαφέρει στη φιλοσοφία της από τη GEO και η επίσημη ιστοσελίδα της είναι: <http://www.ebi.ac.uk/arrayexpress/> [Brazma A et al, 2003].
- **Stanford Microarray Database (SMD)**: δημιουργήθηκε με σκοπό να διαμοιράζονται τα αρχεία μεταξύ ερευνητών του Stanford. Σταδιακά εξελίχθηκε σε δημόσια βάση δεδομένων για μικροσυστοιχίες. Η επίσημη ιστοσελίδα της είναι: <http://smd.princeton.edu/> [Demeter et al., 2007].

2.5 Μετα-Ανάλυση (Meta-Analysis):

Η μετα-ανάλυση είναι μία στατιστική τεχνική συνδυασμού ευρημάτων από ανεξάρτητες μελέτες. Είναι ένα στατιστικό εργαλείο που επεξεργάζεται τα δεδομένα και τα αποτελέσματα μελετών που ερευνούν το ίδιο ερώτημα. Το τελικό αποτέλεσμα που διεξάγει, προέρχεται από τη σύνθεση ανεξάρτητων συνόλων [Normand, 1999].

Για να πραγματοποιηθεί μία μετα-ανάλυση θα πρέπει να καθοριστεί με σαφήνεια το αντικείμενο της μελέτης και κατόπιν να διεξαχθεί αναζήτηση στη βιβλιογραφία για την εύρεση όλων των διαθέσιμων μελετών. Είναι απαραίτητο να συγκεντρωθεί ένας απαιτούμενος αριθμός μελετών για να γίνει η μετα-ανάλυση. Το επόμενο βήμα είναι ο καθορισμός του μεγέθους επίδρασης. Τα μεγέθη επίδρασης είναι τρία και είναι: η διαφορά μέσων τιμών, ο συντελεστής συσχέτισης και ο λόγος αναλογιών OR(Odds ratio).

Όταν τα δεδομένα μας είναι διχοτομικά επιλέγεται σαν μέγεθος επίδρασης ο λόγος αναλογιών OR(odds ratio) ή ο σχετικός κίνδυνος RR(risk ratio) ή η διαφορά κινδύνου RD(risk difference). Ενώ όταν τα δεδομένα είναι συνεχή τότε επιλέγεται η τυποποιημένη διαφορά των μέσων τιμών ή μη τυποποιημένη διαφορά μέσων τιμών μεταξύ των δειγμάτων ελέγχου και αναφοράς.

Η ισχύς και δύναμη μίας μετα-ανάλυσης βασίζεται σε δύο παράγοντες. Ο πρώτος αφορά τον αριθμό μελετών καθώς όσες περισσότερες μελέτες μετα-αναλυθούν τόσο καλύτερα και ισχυρότερα αποτελέσματα εξάγονται. Ο δεύτερος αφορά τη μέθοδο που χρησιμοποιείται για να συνδυαστούν οι μεμονωμένες εκτιμήσεις του μεγέθους επίδρασης οι οποίες προέρχονται από τις αρχικές μελέτες. Δύο μοντέλα χρησιμοποιούνται για το συνδυασμό των παραπάνω εκτιμήσεων :το μοντέλο σταθερών επιδράσεων (fixed effect model) και το μοντέλο των τυχαίων επιδράσεων (random effect model) [Normand, 1999].

Για τα συνεχή δεδομένα, όπως έχει ήδη αναφερθεί, το μέγεθος επίδρασης είναι η διαφορά των μέσων τιμών μεταξύ των δειγμάτων ελέγχου και αναφοράς. Η διαφορά μέσων τιμών μπορεί να είναι είτε τυποποιημένη είτε μη-τυποποιημένη [Thakkestian A, 2005].

Για τον υπολογισμό της μη-τυποποιημένης διαφοράς μέσων τιμών ισχύει ο τύπος:

$$d_i = \bar{x}_{1i} - \bar{x}_{2i}$$

Και για την τυποποιημένη διαφορά μέσων τιμών ισχύει:

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{sd_i}$$

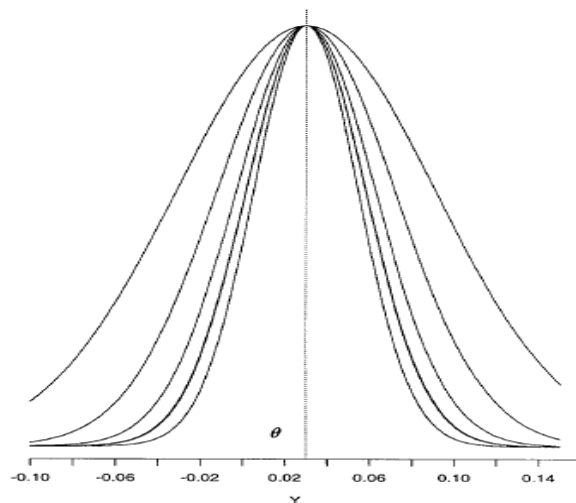
$$sd_i = \sqrt{\frac{(n_{1i} - 1)sd_{1i}^2 + (n_{2i} - 1)sd_{2i}^2}{n_{1i} + n_{2i} - 2}}$$

Όπου: d_i : η τυποποιημένη διαφορά μέσων τιμών και sd_i : η τυπική απόκλιση

2.5.1 Το μοντέλο σταθερών επιδράσεων (fixed effect model):

Το μοντέλο σταθερών επιδράσεων υποθέτει ότι όλα τα δείγματα των μελετών προέρχονται από έναν πληθυσμό που έχουν σαν κοινό μέγεθος επίδρασης έστω θ (παράμετρος). Η κατανομή του μοντέλου ορίζεται ως εξής (Εικόνα 12):

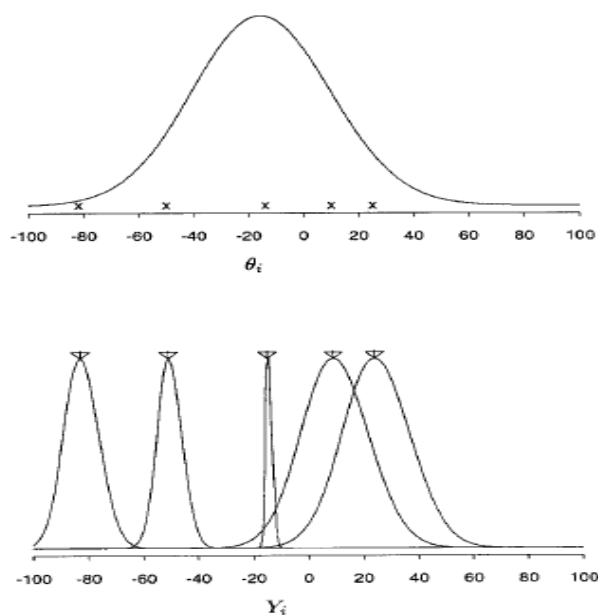
$$Y_i \sim N(\theta, s_i^2), \quad \text{για } i = 1, 2, 3, \dots, k$$



Εικόνα 12: Μοντέλο σταθερών επιδράσεων. Η κατανομή 5 υποθετικών δειγμάτων, πηγή: [Normand, 1999]

2.5.2 Το μοντέλο τυχαίων επιδράσεων (random effect model):

Το μοντέλο αυτό υποθέτει ότι τα δείγματα μίας μελέτης που περιλαμβάνονται στη μετα-ανάλυση, προέρχονται από μία κατανομή πληθυσμών. Σε αυτή τη περίπτωση, κάθε μελέτη έχει διαφορετικό μέγεθος επίδρασης θ_i και διακύμανση s_i^2 με κατανομή: $Y_i | \theta_i, s_i^2 \sim N(\theta_i, s_i^2)$ (Εικόνα 13). Κάθε δείγμα του υπερπληθυσμού έχει μέγεθος επίδρασης το οποίο κατανέμεται με μέση τιμή θ και διακύμανση τ^2 (υπερ-παράμετροι), με κατανομή: $\theta_i | \theta, \tau^2 \sim N(\theta, \tau^2)$.



Εικόνα 13: Μοντέλο τυχαίων επιδράσεων. Η κατανομή 5 υποθετικών δειγμάτων, πηγή: [Normand, 1999]

2.5.3 Προβλήματα κατά τη διεξαγωγή μίας μετα-ανάλυσης:

Για τη διεξαγωγή μίας σωστής μετα-ανάλυσης απαιτείται η εκτενέστερη αναζήτηση για τον εντοπισμό όλων των μελετών οι οποίες είναι διαθέσιμες στη βιβλιογραφία. Για την αποφυγή υπερεκτίμησης των αποτελεσμάτων πρέπει να ληφθούν υπόψη κάποια προβλήματα βιβλιογραφίας όπως είναι η ξενόγλωσση βιβλιογραφία, η γκρίζα βιβλιογραφία, το συστηματικό λάθος δημοσίευσης και το φαινόμενο του Πρωτέα.

Στο φαινόμενο της γκρίζας βιβλιογραφίας εντάσσονται μελέτες οι οποίες δεν έχουν δημοσιευθεί σε κάποιο περιοδικό λόγω των αρνητικών αποτελεσμάτων τους. Στη ξενόγλωσση βιβλιογραφία εντάσσονται μελέτες που διεξάγονται σε μη-αγγλόφωνες χώρες και δημοσιεύονται σε τοπικά περιοδικά με αποτέλεσμα να χάνονται δεδομένα για αυτές τις μελέτες. Τα δύο αυτά προβλήματα αποτελούν το συστηματικό λάθος δημοσίευσης (publication bias) [Egger, 1997].

Επιπλέον ένας ακόμη παράγοντας ο οποίος μπορεί να οδηγήσει σε υπερ/υπό-εκτίμηση των αποτελεσμάτων είναι ο έλεγχος της ετερογένειας μεταξύ των μεμονωμένων μελετών [Thakkinian A, 2005].

Πιο συγκεκριμένα, ειδικά για το μοντέλο τυχαίων επιδράσεων όπου τα δείγματα προέρχονται από διαφορετικούς πληθυσμούς, ο κίνδυνος για ετερογένεια είναι υψηλός. Ωστόσο έχουν αναπτυχθεί διάφορες μέθοδοι (τεστ) που εκτιμούν την ετερογένεια, όπως είναι το Q τεστ του Cochran, ο έλεγχος I^2 και τ^2 . Ανάλογα με τις τιμές των δεικτών, προκύπτει η ύπαρξη ή μη ετερογένειας, παραδείγματος χάριν αν έχει τιμή ίση με τη μονάδα ή πλησιάζει σε αυτήν, υπάρχει ετερογένεια ενώ αν πλησιάζει το μηδέν, δεν υπάρχει.

2.6 Η εφαρμογή της μετα-ανάλυσης στις μικροσυστοιχίες (Meta-analysis in Microarrays):

Όπως έχει ήδη αναφερθεί, η μετα-ανάλυση είναι ένα στατιστικό εργαλείο που επεξεργάζεται τα δεδομένα και τα αποτελέσματα μελετών που ερευνούν το ίδιο ερώτημα. Παρέχει ένα γενικό αποτέλεσμα το οποίο προκύπτει από της σύνθεση των δεδομένων και μελετών. Έτσι και στις μικροσυστοιχίες, όπου γίνεται ταυτόχρονη ανάλυση χιλιάδων γονιδίων προκειμένου να παραχθούν σετ δεδομένων γονιδιακής έκφρασης, χρειάζεται περεταίρω ανάλυση αυτών των προκύπτοντων σετ δεδομένων για να εξαχθεί ένα γενικό-τελικό αποτέλεσμα.

Η εφαρμογή της μετα-ανάλυσης σε δεδομένα μικροσυστοιχιών, είναι αρκετά διαδεδομένη. Αυτό γιατί με τη χρήση της μετα-ανάλυσης είναι δυνατό να εντοπιστούν γονίδια τα οποία εκφράζονται διαφορετικά ανάμεσα φυσιολογικές και παθολογικές καταστάσεις. Για τη διεξαγωγή της, έχουν αναπτυχθεί τρεις τεχνικές που συνδυάζουν και αναλύουν τα γονιδιακές πληροφορίες. Αυτές είναι: μέθοδος των μεγεθών επίδρασης, μέθοδος συνδυασμού των p-value, μέθοδος υπολογισμού του γινομένου των βαθμών κατάταξης (rank product) [Hong, F and R, Breitling, 2008].

2.6.1 Μέθοδος των μεγεθών επίδρασης:

Η συνηθέστερα λεγόμενη ως μέθοδο της διαφοράς μέσω των τιμών. Βασίζεται στη μέθοδο permute για τον υπολογισμό των p-values και την εκτίμηση της τιμής της FDR. Επιλέγεται και χρησιμοποιείται το κατάλληλο μοντέλο μετα-ανάλυσης (μοντέλο σταθερών ή τυχαίων επιδράσεων), υπό την προϋπόθεση ότι τα μεγέθη επίδρασης των μελετών μπορούν να συνδυαστούν.

2.6.2 Μέθοδος συνδυασμού των p-values:

Είναι μια εύχρηστη και κατανοητή μέθοδος καθώς μπορεί να συνδυάσει τις τιμές των p-values από ανεξάρτητες μελέτες. Αυτό προϋποθέτει τη γνώση των τιμών όλων των p-value όλων των γονιδίων και όχι μέρους αυτών. Αν και έχουν αναπτυχθεί διάφορες τεχνικές για το συνδυασμό των p-values, αυτή που είναι πιο κοινή και ευρέως χρησιμοποιούμενη είναι η μέθοδος του αθροίσματος των λογαρίθμων των p-values. Δίνεται από το τύπο:

$$s_i = -2 \sum_{k=1}^K \log(p_{ik})$$

Όπου:

s_i : το μέγεθος επίδρασης, το οποίο είναι το άθροισμα των λογαρίθμων των p-values για κάθε i γονίδιο μεταξύ των μελετών k .

2.6.3 Μέθοδος υπολογισμού του γινομένου των βαθμών κατάταξης (rank product):

Για κάθε γονίδιο υπολογίζεται ο λόγος της έκφρασης μεταξύ του δείγματος ελέγχου και αναφοράς (αλλαγή πτύχωσης). Βάση αυτής της τιμής τα γονίδια κατατάσσονται κατά φθίνουσα σειρά. Έτσι με τη κατάταξη αυτή σε κάθε γονίδιο αντιστοιχεί και ένας αριθμός κατάταξης, ο βαθμός κατάταξής του. Η διαδικασία γίνεται για όλα τα γονίδια όλων των μελετών που συμμετέχουν στη μετα-ανάλυση. Σαν μέγεθος επίδρασης χρησιμοποιείται το γινόμενο των βαθμών κατάταξης το οποίο υπολογίζεται ως εξής:

$$RP_g = (\prod_i \prod_k r_{gik})^{\frac{1}{k}}$$

Επίσης, εναλλακτικός τρόπος είναι υπολογισθεί αντί για το γινόμενο, το άθροισμα ή ο μέσος όρος των βαθμών κατάταξης. Σε κάθε περίπτωση, η στατιστική σημαντικότητα μπορεί να υπολογισθεί μέσω του permutation test [Tseng, 2012].

3 Κεφάλαιο -Υλικά και Μέθοδοι:

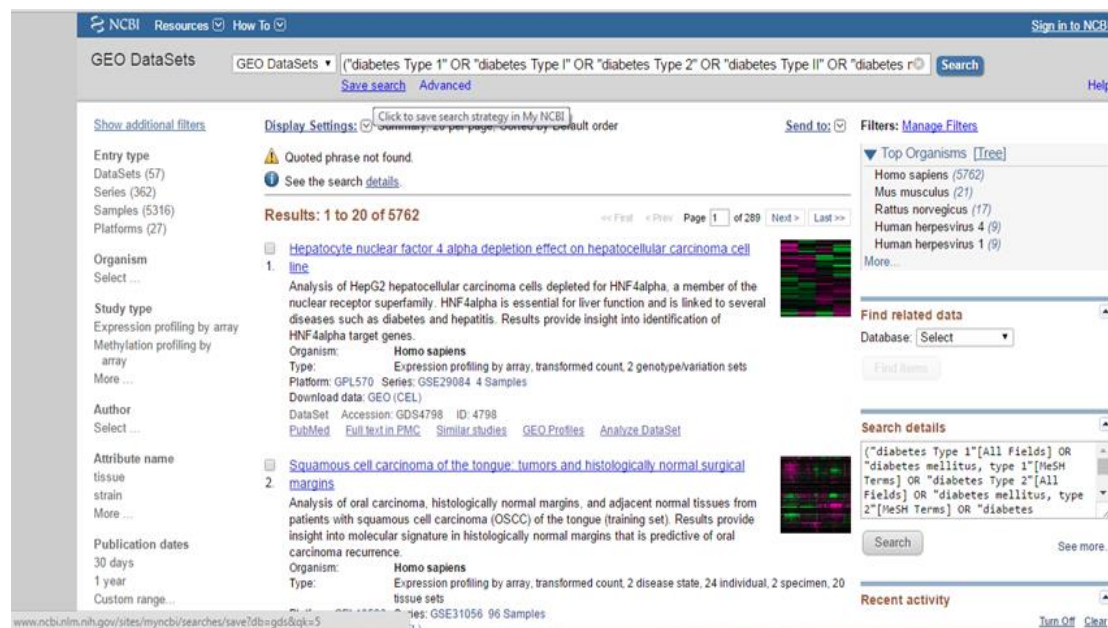
3.1 Ερευνητικό ερώτημα:

Στη παρούσα διπλωματική εργασία, πραγματοποιήθηκε μετα-ανάλυση, με τη χρήση του στατιστικού πακέτου STATA13, σε μικροσυστοιχίες για τη μελέτη της διαφορικής γονιδιακής έκφρασης των γονιδίων της πάθησης του σακχαρώδη διαβήτη. Σκοπός ήταν να εντοπιστούν τα γονίδια τα οποία συσχετίζονται με την εμφάνιση της νόσου.

3.2 Συλλογή και καταγραφή των δεδομένων:

Για τον εντοπισμό και τη συλλογή των δεδομένων, πραγματοποιήθηκε αναζήτηση στη βάση δεδομένων της GEO (Gene Expression Omnibus). Η GEO αποτελεί δημόσιο αποθετήριο δεδομένων γονιδιακής έκφρασης, όπου η υποβολή των δεδομένων γίνεται σύμφωνα με το πρωτόκολλο του MIAME (Minimum Information About Microarray Experiment). Για την ανάκτηση των σετ δεδομένων δίνεται η δυνατότητα διατύπωσης ενός ερωτήματος (query).

Στη προκειμένη περίπτωση, όπου ελέγχεται η διαφορική έκφραση των γονιδίων σχετικά με το σακχαρώδη διαβήτη παρουσιάζεται ένα απόσπασμα αποτελέσματος μετά την εφαρμογή του ερωτήματος (Εικόνα 14):



Εικόνα 14: Η βάση δεδομένων γονιδιακής έκφρασης GEO του NCBI (κατά την εφαρμογή του ερωτήματος) .

Κατά την εισαγωγή του ερωτήματος στη GEO Datasets, εμφανίστηκαν όλα τα σχετικά αποτελέσματα-σετ δεδομένων. Ωστόσο, περιορίστηκε η αναζήτησή στο πεδίο των series (πάνω αριστερά στην Εικόνα 14: Entry type: Datasets, Series, Samples, Platform) όπου προέκυψαν λιγότερες μελέτες. Έτσι το εύρος των δεδομένων που δόθηκε ήταν πλέον μικρότερο.

Για κάθε μελέτη, εξετάστηκε η περίληψη της (summary) και αν δεν ήταν εμφανές αν πρόκειται για κάποια που θα χρησιμοποιηθεί, αναζητήθηκε το αντίστοιχο άρθρο στη βάση δεδομένων της Pubmed, μέσω του κωδικού PMID.

Αν και οι μελέτες που προέκυψαν, κατόπιν του περιορισμού τους με την επιλογή Series, αναφερόντουσαν σε σακχαρώδη διαβήτη ωστόσο δεν χρησιμοποιήθηκαν όλες στη μετ-ανάλυση. Ο λόγος απόρριψής τους ήταν είτε διότι εξέταζαν την έκφραση της νόσου σε διάφορα στάδια της, είτε διότι αναφερόντουσαν σε τυχόν θεραπείες φαρμάκων που πραγματοποιούνταν σε διαβητικούς ασθενείς, είτε γιατί απλά αναφερόταν ως ένας παράγοντας προδιάθεσης στην εμφάνιση κάποιας άλλης νόσου. Άλλες απορριφθήκαν για λόγους του ότι εξέταζαν micro RNA οπότε ήταν εκτός πεδίου της συγκεκριμένης έρευνας και άλλες δεν περιείχαν ασθενείς και μάρτυρες ώστε να μπορέσουν να ομαδοποιηθούν.

Για τις μελέτες εκείνες που τελικά χρησιμοποιήθηκαν, ανακτήθηκε ο πίνακας δεδομένων γονιδιακής έκφρασης (series matrix file) και τα δεδομένα του αρχείου της πλατφόρμας (GPL). Στο πρώτο πίνακα (series matrix file) περιέχονταν τα probes με τις τιμές έκφρασης για κάθε δείγμα (χαρακτηριστικά όπως Id, Id_ref) και στη πλατφόρμα, η ονομασία των γονιδίων που χρησιμοποιήθηκαν (αλλά και άλλα χαρακτηριστικά όπως GB_ACC, SPOT_ID, Id).

Με την ανάκτηση και την αντιπαραβολή αυτών των 2 πινάκων, μέσω του πεδίου Id, ταυτοποιήθηκαν οι ονομασίες των γονιδίων. Για λόγους συμβατότητας των ονομάτων των γονιδίων, πραγματοποιήθηκε αναζήτηση στη βάση δεδομένων [HGNC HUGO](#), η οποία είναι υπεύθυνη για την έγκριση μοναδικών συμβόλων και ονομάτων για τις τοποθεσίες γονιδίων και τα γονίδια του ανθρώπου, συμπεριλαμβανομένων των γονιδίων που κωδικοποιούν πρωτεΐνες, των γονιδίων lncRNA και των ψευδο-γονιδίων, επιτρέποντας σαφή επιστημονική επικοινωνία.

Έτσι, αφού έγιναν οι απαραίτητες αντικαταστάσεις των ονομάτων των γονιδίων με το καθολικό συμβολισμό, βάσει της HUGO, δημιουργήθηκε ένα αρχείο excel με τα εξής χαρακτηριστικά: Γονίδιο (Gene/ GeneSymbol) και τα δείγματα ασθενών και μαρτύρων με τις τιμές έκφρασής τους (cases/ controls) (Εικόνα 15).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GeneSymbol	GSM228562	GSM228563	GSM228564	GSM228565	GSM228566	GSM228567	GSM228568	GSM228569	GSM228570	GSM228571	GSM228572	GSM228573
2	A1CF	171	210.7	194.9	169.4	237.1	144.5	228.5	118.2	174.8	220.3	375.8	177.2
3	A2M	17.6	10.8	6.7		11 14.4		7 15.4		64 56.8	36.5		97 86.9
4	A4GALT		79 25.5	44.9	16.9	34.9	47.1	17.7	38.3	37.7	61.6	110.8	54.3
5	A4GNT	99.2	87.8	128.9		48 143.8	171.8	171.9	77.2	145.8	147.2	255.7	118.1
6	AAAS	299.8	186.5	119.7	188.5	230.5	98.3	41.6	134.1	138.3	165.5	125.6	163.2
7	AACS	200.9		233 230.8	221.7	300.2	216.5	415.5	222.3		267 252.2		82 333.9
8	AADAC	10.4	19.5	9.9	8.6	17.3	6.6		17 16.6	18.1		15 128.3	57.9
9	AAGAB	189.15	160.6	207.35	216.15	122.85	198.6	273.75	256.75		168 175.85	140.1	260.45
10	AAK1	342.06	288.26	242.12	259.14	288.32	223.76	256.36	265.24	450.02	670.34	511.16	339.86
11	AAMP		595 817.5	903.7	716.9	614.4	881.3	717.6	563.7	908.5	906.1	512.5	652.4
12	AANAT	13.7	10.9	10.7	10.7	7.3	3.9	4.5	6.5	10.6	6.7	13.9	9.5
13	AARS	1049.1	1032.5	1131.2	1125.9	1199.6	1083.4	1022.3	1304.8	1011.2	1112.2	982.4	1335.3
14	AARSD1		332 238.5	312.4	293.7	391.8	307.7	312.4	314.5	422.2	315.6	417.3	250.9
15	AASDHPPT	601.25	387.75	561.15	661.4	366.1	671.65	500.7	858.65	376.8	523.55	404.75	518.2
16	AASS		41 19.05	20.35	17.7	57.1	14.85	45.45	41.2	26.5	9.2	16.35	19.15
17	AATF	933.2	1293.2	894.5		977 1571.3	1165.8	970.8	1089.2	1292.6	802.2	778.9	704.1
18	AATK	355.1	349.8	351.9	221.4	188.7	317.8	223.5	253.1	273.9	159.2		320 380.7

Εικόνα 15: Απόσπασμα αρχείου excel της μελέτης GSE9006.

Ωστόσο, αν σε μία εγγραφή κάποιου ανιχνευτή δεν αντιστοιχίζοταν κάποιο γονίδιο λόγω μη υβριδοποίησης ή λόγω dark-bright corner ή υπήρχε συνδυασμός γονιδίων (εξέταση 2,3ή και 4 γονιδίων μαζί), αυτή διαγραφόταν. Επιπρόσθετα, στη στήλη των γονιδίων υπήρχαν πολλά (γονίδια) με την ίδια ονομασία αλλά με διαφορετικές τιμές έκφρασης, και

αυτό γιατί πολλοί ανιχνευτές υβριδοποιούνταν από πολλαπλά μετάγραφα του ίδιου γονιδίου. Αυτές οι εγγραφές αντικαταστάθηκαν από μία, η οποία είχε τιμές έκφρασης τις μέσες τιμές των εγγραφών αυτών. Η υλοποίηση αυτής της αντικατάστασης έγινε στο στατιστικό πακέτο του STATA13 με την εντολή `collapse`. Και πιο συγκεκριμένα (για τη μελέτη GSE9006) :

`collapse (mean) GSM228562-GSM228666,by(GeneSymbol)`

Με τον τρόπο αυτό επιτεύχθηκε η μετατροπή των probes σε γονίδια ή περιοχές γονιδίων. Η διαδικασία έγινε για όλες τις μελέτες που χρησιμοποιήθηκαν στη παρούσα εργασία. Όταν τελικά πραγματοποιήθηκαν και οι απαραίτητες διορθώσεις, οι πίνακες των εκάστοτε μελετών ήταν έτοιμοι για επεξεργασία.

3.3 Στατιστική ανάλυση των δεδομένων:

Καθώς, οι μελέτες είχαν διαφορετική βαρύτητα στο μέγεθος επίδρασης και δεν είχαν την ίδια επίδραση στο πληθυσμό, επιλέχθηκε το μοντέλο τυχαίων επιδράσεων. Σαν μέγεθος επίδρασης χρησιμοποιήθηκε η τυποποιημένη διαφορά μέσων τιμών των ασθενών και μαρτύρων (cases/controls). Έτσι με εφαρμογή του t τεστ, υπολογίστηκε η διαφορά t και το τυπικό σφάλμα (standard error). Οι τύποι υπολογισμοί των παραπάνω δίνονται ως εξής:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}$$

Όπου:

$\bar{x}_1 - \bar{x}_2$: η διαφορά μέσων τιμών του δείγματος

$\mu_1 - \mu_2$: η διαφορά μέσων τιμών του πληθυσμού

$S_{\bar{x}_1 - \bar{x}_2}$: το τυπικό σφάλμα

Στις μικροσυστοιχίες συνηθίζεται να υπάρχει μικρό μέγεθος δείγματος και παράλληλα οι τιμές των τιμών έκφρασης έχουν μη κανονική κατανομή, έτσι εκτελείται μετά το t test, η μέθοδος bootstrap. Με τη μέθοδο αυτή πετυχαίνεται η ταυτοποίηση των διαφορετικά εκφρασμένων γονιδίων.

Η αρχική ενέργεια ήταν στη τελευταία γραμμή να προστεθούν ακόμη άλλες 2. Η μία ήταν συμπληρωμένη με μηδέν και ένα, με παραδοχή ότι το 0 αντιστοιχεί στους υγιείς και το 1 στους ασθενείς και η άλλη ήταν συμπληρωμένη με τον αριθμό της κάθε μελέτης, για παράδειγμα αν υπήρχαν 4 μελέτες κάθε μελέτη θα είχε έναν αριθμό από το 1 έως και το 4

Στη συνέχεια ακολούθησε ο έλεγχος t test για κάθε μελέτη (GSE) με σκοπό να εντοπιστούν τα γονίδια που εκφράζονταν διαφορετικά. Ο έλεγχος πραγματοποιήθηκε δίνοντας των εντολή:

`ttest `var', by(case_control) uneq`

και κατόπιν ακολούθησε η μέθοδος bootstrap:

`bootstrap t=r(t), reps(100) strata(case_control): ttest `var',by(case_control) uneq`

Με τη τελευταία εντολή επανα-υπολογίστηκαν νέα διαστήματα εμπιστοσύνης, νέες τιμές του τυπικού σφάλματος και νέες τιμές του p value για κάθε γονίδιο. Όλες οι εντολές έγιναν για όλες τις μελέτες που χρησιμοποιήθηκαν στην εργασία. Ένα απόσπασμα από το

αποτέλεσμα των παραπάνω ελέγχων στη μελέτη GSE9006 του σακχαρώδη διαβήτη τύπου 1 παρουσιάζεται στην Εικόνα 16.

gene	t	r(se)	p	e(se)
a1cf	-.0399189070194768	16.11794614658873	.9682940966310475	1.044184350923976
a2m	.3400936198892486	8.423639197562935	.7351468201926239	1.175980293782597
a4galt	.5582282518909885	10.59102450132388	.5797697456424678	.9876914955284808
a4gnt	1.33902419015135	14.25073046767726	.1868451442436689	.8784896782189025
aaaa	.3442792654733804	16.97879022106685	.7321753586337825	.9735018004851003
aacs	.2491642251274361	18.4418784600215	.8044042592991909	1.100916302875589
aadac	1.540309912006397	6.643621759803887	.1349308233818966	.7186052581308455
aagab	.3164101577656093	12.45899248997107	.753172697894114	.9877107428921808
aak1	1.624821009874758	32.85952820192477	.1140186224519462	.9230455945196492
aamd	-1.688065715101323	5853983.853542891	.0962465800592359	.9204860260119675
aamp	-.9163374915864264	39.67079470326999	.3633658082174663	1.115232502436467
aanat	.0525061890793934	1.057461105784264	.9582984719058627	.9198501341726826
aar2	-.8783207941528988	21.86629637663946	.3832456442733156	.9516309596909354
aars	-2.140204484954198	60.68272788709543	.0368645480308367	.9580276221529527
aarsd1	-.9987887141217956	15.82032346585341	.3219416460064026	1.07427966520001
aarsdpp	-.8932267541262149	53.23737468398253	.3756731034858141	1.027853662432022
aasa	1.781322493419093	3.30251748577992	.0818373967676096	.8620044058687872
aatf	.77246976666087	62.29792199495569	.4450781832206076	.9695397367162029
aatx	-3.309098931513583	26.14473220113901	.0015798764604494	.9961574514804961
abat	.5126367922665244	3733234.216766466	.6113355251987245	.986827901617277
abca1	-1.068169836639336	3048410.847747533	.2895222634864645	.9782968148336385
abcallp	-2.688640197975978	7.77315782711853	.0094719953868899	1.195640274162479
abca12	1.644667565293174	6.436999735393705	.1074264558655697	.9693118318505566
abca2	-.4862147916466144	8374234.359376552	.6291084310089283	1.026403235791167
abca3	-.6709221239817903	3.987916898927541	.5047237954884118	1.017164422113764
abca4	.3593362848263556	4.253373532144846	.7210473175827474	.8974056010243164
abca5	-.4822164935523824	25.19615171086861	.6323000461062136	1.160056564724838
abca6	1.209234238455657	5.671232992930939	.2324461700572167	1.098014382460877
abca7	-2.241061496603361	23.16312063755819	.0293671654464333	.9204743735943893
abca8	.5850897088754256	5.171969435625219	.5612913431485394	1.076632944084345
abcb1	-.0941570028464145	17.63917222575843	.9253340209390797	1.024496826367254
abcb11	-1.092992583035694	4.444495354521616	.2789498717414088	1.036126412631705
abrch4	-.6378744682775526	19.23645615529791	.5271127358824976	1.085914690843341

Εικόνα 16: Απόσπασμα αποτελέσματος t-test & bootstrap στη μελέτη GSE9006.

Τα αρχεία κειμένου τα οποία προέκυψαν ανά μελέτη (GSE) ενώθηκαν -ξεχωριστά για κάθε τύπο διαβήτη- όλα μαζί προκειμένου να διεξαχθεί η μετα-ανάλυση. Με τη χρήση της εντολής *append using* έγινε οι σύνθεση των μελετών ενώ με τη παρακάτω εντολή πραγματοποιήθηκε η μετα-ανάλυση:

```
metan t rse if ngene==`i',nograph randomi
```

Ο ολοκληρωμένος κώδικας που χρησιμοποιήθηκε για τους παραπάνω ελέγχους παραθέτεται στο Παράρτημα 2. Το αποτέλεσμα της μετα-ανάλυσης είναι ένα αρχείο κειμένου στο οποίο προστέθηκε μία ακόμη στήλη, η ονομασία των εκάστοτε γονιδίων. Στην Εικόνα 17 παρουσιάζεται ένα απόσπασμα αποτελέσματος της μετα-ανάλυσης. Αυτό που ελέγχεται είναι η τιμή του p-value προκειμένου να εντοπιστούν τα στατιστικώς σημαντικά γονίδια.

Τα αποτελέσματα της μετα-ανάλυσης δίνουν p-values τα οποία εντοπίζουν λανθασμένα στατιστικώς θετικά γονίδια (false positive). Τα επίπεδα σημαντικότητας (5% και 1%) περιέχουν λανθασμένα στατιστικώς σημαντικά γονίδια, καθώς ο αριθμός των γονιδίων που εξετάζονται στις μικροσυστοιχίες είναι πολύ μεγάλος. Προκειμένου να εντοπιστούν (τα στατιστικώς σημαντικά γονίδια) εφαρμόστηκαν οι μέθοδοι διόρθωσης του p-value (γίνεται αναφορά για αυτές στο Κεφάλαιο 2).

finalresults2.txt											
		1	2	3	4	5	6	7	8	9	10
		gene	ES	se(ES)	z	p	df				
1	2	.7417214080381328	1.056662566156776	.7019472741765364	.4827120456154056	1					
2	3	-.381986970961562	1.304484530965704	.2928259875023422	.7696551633474209	0					
3	4	-.1614354272562574	.7670065280383357	.2104746457232091	.833297238234576	1					
4	5	.0382629470177797	1.55233066451192	.0246487091265508	.9803351668216412	0					
5	6	1.147075415004638	1.306680676350729	.8778544259246006	.3800227240717862	0					
6	7	-.3412727113540711	.6473207197509233	.5272080762151199	.5980491008681327	2					
7	8	1.086239501336375	.7978619873771798	1.361437840781438	.1733753645203838	1					
8	9	.2086163363215635	.7216514629806279	.2890818449392704	.7725187464038535	1					
9	10	.3743453695252733	.701351252742579	.5337487714770952	.5935153610929302	2					
10	11	1.396745681329842	9.212733014494983	.1516103504934153	.8794942692721068	0					
11	12	.0924734942874606	1.061542514753485	.0871123793934293	.9305821854613121	0					
12	13	1.323361758667913	.6923567599123478	1.911387069919633	.0559548570976044	2					
13	14	-.6792735090919688	1.478282101160836	.4595019506483657	.6458737515935352	2					
14	15	.0978914814727061	1.160815838539903	.0843298981825024	.9327941417486739	0					
15	16	-.9021980545342165	.7034630642368026	1.282509488274305	.1996639786657335	2					
16	17	.0774387967032319	1.253510231322734	.0617775545569473	.950739978207648	0					
17	18	-1.620725779139748	.7569882866618175	2.14101830595934	.0322725618545644	1					
18	19	1.477607989402477	.7613568280136042	1.940756206597092	.0522878552254512	1					
19	20	.8230485704091419	1.527816568659883	.538709022596263	.5900876470806185	1					
20	21	.3131729844732969	.9483855359547004	.3302169556582689	.741236035999282	2					
21	22	.337364607699067	1.080260096180783	.3122994257510819	.7548129764271658	2					
22	23	-.1569499978943254	.915496730367644	.171436983539305	.8638801765591089	2					
23	24	-1.383182402055843	1.1890669981715	1.16325018201905	.2447280117000768	2					
24	25	.6602369321189934	.9524184548701412	.693221481317279	.4881705693380192	0					
25	26	.948757032592765	.5725844916586168	1.656972982003899	.0975249376105219	2					
26	27	.8267721254768388	8.493951681649753	.0973365703578217	.9224591149258971	0					
27	28	.20174993985015	1.044828157361799	.1930938962820169	.8468854327660338	1					
28	29	-.6245499653480739	.7418561800030441	.8418747220593501	.3998580808937332	2					

Εικόνα 17: Απόσπασμα αποτελέσματος μετα-ανάλυσης.

Ειδικότερα με τη χρήση των εντολών:

```
multproc, puncor(0.05) pval(p) meth(simes) rej(fdr)
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)
```

πραγματοποιήθηκε η εύρεσή τους. Η πρώτη εντολή αφορά τη μέθοδο διόρθωσης κατά FDR τη μία φορά με $p=0.05$ και τη δεύτερη με $p=0.01$. Η τρίτη αφορά τη μέθοδο bonferroni, η τέταρτη τη μέθοδο sidak, η πέμπτη τη holm και η τελευταία τη Holland. Το αποτέλεσμα της κάθε μεθόδου ήταν το 0 και το 1. Με 1 εμφανίζονταν τα στατιστικά σημαντικά γονίδια. Η πρώτη μέθοδος, που είναι και πιο συντηρητική (FDR), είχε σαν αποτέλεσμα μεγαλύτερο αριθμό στατιστικά σημαντικών γονιδίων σε σχέση με τις υπόλοιπες. Ενώ από κοινού όλες μαζί έβγαζαν τα ίδια στατιστικά σημαντικά γονίδια. Τα αποτελέσματα των εντολών ήταν 0 και 1, με συμφωνία ότι με 1 εμφανίζονται τα στατιστικώς σημαντικά γονίδια

3.4 Η χρήση των πλατφορμών bioCompedium, Panther και STRING :

Η bioCompedium είναι μία προσβάσιμη από το κοινό, υψηλής απόδοσης πλατφόρμα πειραματικής ανάλυσης δεδομένων. Το σύστημα είναι σχεδιασμένο να λειτουργεί με μεγάλες λίστες γονιδίων ή πρωτεϊνών για τις οποίες συλλέγει ένα ευρύ φάσμα βιολογικών πληροφοριών. Διευκολύνει την ανάλυση, σύγκριση και τον εμπλουτισμό των πειραματικών αποτελεσμάτων είτε ιδιόκτητα είτε διαθέσιμα στο κοινό σύνολα δεδομένων. Τυπικές περιπτώσεις χρήσης είναι η ιεράρχηση των πιθανών στόχων από μελέτες ανάλυσης γονιδιακής έκφρασης ή από μελέτες RNAi.

Η πλατφόρμα της PANTHER (Protein ANalysis THrough Evolutionary Relationships) αποτελεί σύστημα ταξινόμησης και σχεδιάστηκε για να ταξινομεί τις πρωτεΐνες (και τα γονίδια τους), προκειμένου να διευκολυνθεί η ανάλυση υψηλής απόδοσης. Οι πρωτεΐνες ταξινομούνται σύμφωνα με:

- Οικογένεια και υποοικογένεια: οικογένειες είναι ομάδες εξελικτικών πρωτεϊνών που σχετίζονται και οι υποοικογένειες αναφέρονται σε πρωτεΐνες που έχουν επίσης την ίδια λειτουργία.
- Μοριακή λειτουργία: η λειτουργία της πρωτεΐνης ή η αλληλεπίδρασή της με πρωτεΐνες σε ένα βιοχημικό επίπεδο, π.χ. μια πρωτεϊνική κίνηση.
- Βιολογική διαδικασία: η λειτουργία της πρωτεΐνης στα πλαίσια ενός μεγαλύτερου δικτύου πρωτεϊνών που αλληλεπιδρούν για να επιτευχθεί μια διαδικασία στο επίπεδο του κυττάρου ή του οργανισμού, π.χ. μίτωση.
- Βιοχημικά/μεταβολικά μονοπάτια: παρόμοια με τη βιολογική διαδικασία, αλλά ένα μονοπάτι καθορίζει ρητά τις σχέσεις μεταξύ των αλληλεπιδρώντων μορίων.

Η βάση δεδομένων STRING αποτελεί βιολογική βάση και διαδικτυακή πηγή των γνωστών και των προβλεπόμενων αλληλεπιδράσεων πρωτεΐνης με πρωτεΐνη.

Στα εργαλεία της bioCompedium, της Panther και της STRING, τοποθετήθηκαν οι λίστες με τα στατιστικώς σημαντικά γονίδια που προέκυψαν μετά την εφαρμογή των μεθόδων διόρθωσης, προκειμένου να συλλάβουν τα μοριακά και βιοχημικά χαρακτηριστικά των επιλεγμένων γονιδίων αλλά και να αναπαρασταθούν μέσω δικτύου οι τυχόν αλληλεπιδράσεις τους. Πιο συγκεκριμένα τα στατιστικώς σημαντικά γονίδια που προέκυψαν από τη μέθοδο κατά FDR αποτέλεσαν την είσοδο στις πλατφόρμες.

Για κάθε τύπο διαβήτη δημιουργήθηκαν 2 αρχεία excel όπου καταχωρήθηκαν τα χαρακτηριστικά της κάθε μελέτης. Οι μελέτες που εν τέλει χρησιμοποιήθηκαν παρουσιάζονται στην Εικόνα 19.

Type 1 Diabetes Mellitus					No of probes	No of genes
GSE	No.	cases	controls	Platform		
GSE9006	1	43	24	GPL96[HG-U133A] Affymetrix Human Genome U133A Array	22283	12749
GSE55098	2	12	10	GPL570[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	54675	20262
GSE29142	3	9	10	GPL13507:Phalanx Human OneArray (version 4.3)	29187	13867
GSE30210	4	18	18	GPL6947:illumina HumanHT-12 V3.0 expression beadchip	49576	24756
GSE17635	5	11	11	GPL2700:Sentrix HumanRef-8 Expression BeadChip	24354	12738
GSE33440	6	16	6	GPL6947:illumina HumanHT-12 V3.0 expression beadchip	48803	24909
Type 2 Diabetes Mellitus						
GSE	No.	cases	controls	Platform		
GSE4117	1	4	4	GPL1708:Agilent-012391 Whole Human Genome Oligo Microarray G4112A (Feature Number version)	41675	17831
GSE15932	2	8	8	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	54675	20262
GSE9006	3	12	24	GPL96[HG-U133A] Affymetrix Human Genome U133A Array	22283	12749

Εικόνα 19: Συγκεντρωτικός πίνακας των μελετών (GSE) με τον αντίστοιχο αριθμό κάθε μελέτης, τους ασθενείς και μάρτυρες, τη πλατφόρμα που χρησιμοποιήθηκε, τον αριθμό των ανιχνευτών και των γονιδίων.

Για κάθε μελέτη ανακτήθηκε ο πίνακας γονιδιακής έκφρασης (series matrix file) και το αντίστοιχο αρχείο της πλατφόρμας (GPL). Οι ανιχνευτές που δεν είχαν ονομασία ή δεν περιείχαν δεδομένα (τιμή NULL) όπως επίσης πολλά γονίδια που ελεγχόντουσαν μαζί με κάποια άλλα διαγράφηκαν. Παράλληλα, λόγω του ότι πολλοί ανιχνευτές αναφερόντουσαν στο ίδιο γονίδιο, προέκυπταν πολλαπλά καταχωρημένα γονίδια με διαφορετικές τιμές έκφρασης. Για την αποφυγή αυτών των πολλαπλών εγγραφών πραγματοποιήθηκε η αντικατάστασή τους με μία εγγραφή με τον αντίστοιχο μέσο όρο τους στις τιμές έκφρασης (μετατροπή probes σε γονίδια). Η εντολή collapse (mean) GSE....by(GeneSymbol) του STATA13 ήταν αυτή που έδωσε την επιθυμητή αντικατάσταση.

4.2 Δεδομένα προς στατιστική ανάλυση και μετα-ανάλυση :

Η ύπαρξη ετερογένειας στα δεδομένα μεταξύ των δειγμάτων οδήγησε στη χρήση του μοντέλου τυχαίων επιδράσεων. Για καλύτερα αποτελέσματα, σαν μέγεθος επίδρασης χρησιμοποιήθηκε η τυποποιημένη διαφορά μέσων τιμών. Η κατανομή της μέσης τιμής εκτιμήθηκε με τον έλεγχο t-test για κάθε μελέτη ξεχωριστά. Προκειμένου να αποφευχθεί το πρόβλημα του μικρού μεγέθους δείγματος και μη κανονικής κατανομής των δεδομένων, έγινε διόρθωση μέσω της μεθόδου bootstrap (η οποία πραγματοποιεί επαναδειγματοληψία) όπου υπολογίστηκαν νέα διαστήματα εμπιστοσύνης και τυπικό σφάλμα με μεγαλύτερη ακρίβεια. Τα αποτελέσματα που προέκυψαν (για κάθε τύπο διαβήτη ξεχωριστά), τα οποία ήταν αρχεία κειμένου (Εικόνα 16) που περιείχαν τιμές του p-value, ενώθηκαν για να κατευθυνθούν προς τη διαδικασία της μετα-ανάλυσης.

Ο αριθμός των γονιδίων κάθε μελέτης (GSE) παρουσιάζεται στους πίνακες 1 και 2.

Πίνακας 1: Αριθμός γονιδίων ανά μελέτη για το διαβήτη τύπου 1

Type 1 Diabetes						
GSE	GSE9006	GSE17635	GSE29142	GSE30210	GSE33440	GSE55098
No of genes	12754	16739	13876	24997	25159	20307

Πίνακας 2: Αριθμός γονιδίων ανά μελέτη για το διαβήτη τύπου 2

Type 2 Diabetes			
GSE	GSE4117	GSE9006	GSE15932
No of genes	18832	12754	20307

Τα αρχεία που δημιουργήθηκαν με το πέρας της μετα-ανάλυσης, τα οποία ήταν και πάλι αρχεία κειμένου, περιείχαν τις τιμές του τυπικού σφάλματος, του p-value και του z για κάθε γονίδιο (Εικόνα 20).

gene	es	sees	z	p	df
a5	1	.741721	1.05666	.701947	.482712
a1bg	2	-.381987	1.30448	.292826	.769655
a2bp1	3	-.161435	.767007	.210475	.833297
a2m	4	.038263	1.55233	.024649	.980335
a2m1	5	1.14708	1.30668	.877854	.380023
a4gnt	6	-.341273	.647321	.527208	.598049
aaas	7	1.08624	.797862	1.36144	.173375
aacs	8	.208616	.721651	.289082	.772519
aadac	9	.374345	.701351	.533749	.593515
aadac11	10	1.39675	9.21273	.15161	.879494
aadac12	11	.092473	1.06154	.087112	.930582
aadat	12	1.32336	.692357	1.91139	.055955
aak1	13	-.679273	1.47828	.459502	.645874
aamp	14	.097891	1.16082	.08433	.932794
aanat	15	-.902198	.703463	1.28251	.199664
aars	16	.077439	1.25351	.061778	.95074
aarsd1	17	-1.62073	.756988	2.14102	.032273
aars1	18	1.47761	.761357	1.94076	.052288
aasdh	19	.823049	1.52782	.538709	.590088
aasdhppt	20	.313173	.948386	.330217	.741236
aass	21	.337365	1.08026	.312299	.754813
aatf	22	-.15695	.915497	.171437	.86388
aatk	23	-1.38318	1.18907	1.16325	.244728
abat	24	.660237	.952418	.693222	.488171
abc1	25	.948757	.572585	1.65697	.097525
abca1	26	.826772	8.49395	.097337	.922459
abca10	27	.20175	1.04483	.193094	.846885
abca11	28	-.62455	.741856	.841875	.399858

Εικόνα 20: Απόσπασμα αποτελέσματος μετα-ανάλυσης σε dta μορφή μετά τη προσθήκη της στήλης με την ονομασία των γονιδίων(data editor STATA13).

4.3 Εύρεση των στατιστικώς σημαντικών γονιδίων:

Όπως ήδη έχει προαναφερθεί, τα επίπεδα σημαντικότητας 5% και 1% ($p\text{-value}=0.05$ & $p\text{-value}=0.01$ αντιστοίχως), έχουν σαν αποτέλεσμα την ύπαρξη λανθασμένων στατιστικώς σημαντικών γονιδίων. Προς αποφυγή αυτού, πραγματοποιήθηκαν οι μέθοδοι διόρθωσης του p-value και συγκεκριμένα οι Simes (FDR) για $p=0.01$ και για $p=0.05$, Bonferroni, Holm, Sidak και Holand.

Τα στατιστικά σημαντικά γονίδια που προέκυψαν μετά τον έλεγχο t-test&bootstrap αλλά και μετά της εφαρμογής της μεθόδου διόρθωσης FDR για κάθε μελέτη για τους δύο τύπους διαβήτη, παρουσιάζονται συνολικά στους παρακάτω πίνακες (Πίνακας 3&4).

Πίνακας 3: Ο αριθμός των στατιστικά σημαντικών γονιδίων για τα επίπεδα σημαντικότητας 1% και 5% και για τη μέθοδο FDR για κάθε μελέτη του Διαβήτη τύπου 1.

Type 1 Diabetes						
GSE	GSE9006	GSE17635	GSE29142	GSE30210	GSE33440	GSE55098
p<0.05	1348	728	796	75	1456	786
p<0.01	508	176	168	2	228	106
FDR	159	17	0	0	1	1

Πίνακας 4: Ο αριθμός των στατιστικά σημαντικών γονιδίων για τα επίπεδα σημαντικότητας 1% και 5% και για τη μέθοδο FDR για κάθε μελέτη του Διαβήτη τύπου 2.

Type 2 Diabetes			
GSE	GSE4117	GSE9006	GSE15932
p<0.05	534	2022	1940
p<0.01	76	932	591
FDR	0	145	3

Η εφαρμογή των μεθόδων διόρθωσης πραγματοποιήθηκε για κάθε μελέτη χωριστά προκειμένου να συγκριθούν τα στατιστικά σημαντικά γονίδια κάθε μίας με αυτά της μετα-ανάλυσης. Κατόπιν της σύγκρισης διαπιστώθηκε ότι κανένα γονίδιο από αυτά της μετα-ανάλυσης (μετά την εφαρμογή της μεθόδου FDR) δεν εντοπίστηκε στα αποτελέσματα των t-test & bootstrap κάθε μελέτης. Ωστόσο κατά τον έλεγχο με επίπεδο σημαντικότητας 5% , προέκυψαν κοινά γονίδια μεταξύ τους. Στο Πίνακα 5 και 6 παρουσιάζεται ο αριθμός των στατιστικά σημαντικών γονιδίων που εντοπίστηκαν και στη μετα-ανάλυση αλλά και κατά τη διεξαγωγή των δύο ελέγχων (κάθε μελέτης).

Πίνακας 5: Ο αριθμός των κοινών στατιστικά σημαντικών γονιδίων μεταξύ μετα-ανάλυσης και των ελέγχων t-test & bootstrap για τις μελέτες του διαβήτη τύπου 1

Type 1 Diabetes						
GSE	GSE9006	GSE17635	GSE29142	GSE30210	GSE33440	GSE55098
meta-analysis	90	37	47	6	84	43

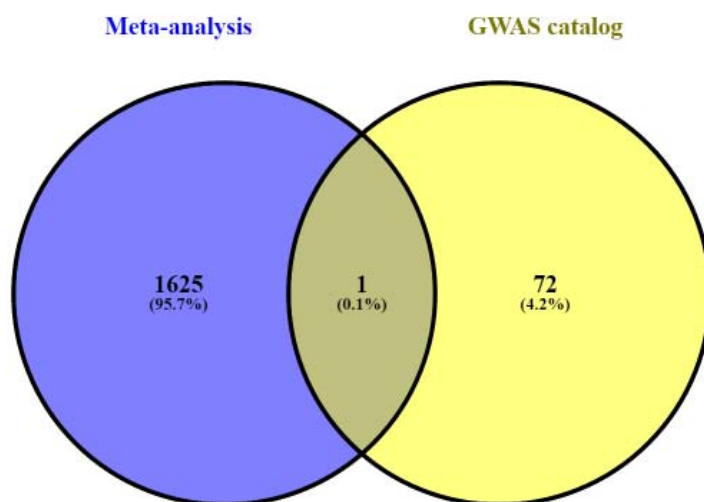
Πίνακας 6: Ο αριθμός των κοινών στατιστικά σημαντικών γονιδίων μεταξύ μετα-ανάλυσης και των ελέγχων t-test & bootstrap για τις μελέτες του διαβήτη τύπου 1

Type 2 Diabetes			
	GSE4117	GSE9006	GSE15932
meta-analysis	38	33	127

Μετά την εφαρμογή των μεθόδων διόρθωσης στα αποτελέσματα της μετα-ανάλυσης, ο αριθμός των στατιστικά σημαντικών γονιδίων περιορίστηκε αρκετά. Χαρακτηριστικό όλων των αποτελεσμάτων ήταν ο κοινός εντοπισμός των γονιδίων σε όλες τις μεθόδους διόρθωσης και το γεγονός ότι όλες οι μέθοδοι εντόπιζαν ως στατιστικώς σημαντικά γονίδια αυτά με το μικρότερο p-value. Στο Πίνακα 7 παρουσιάζονται αναλυτικά τα αποτελέσματα για κάθε τύπο διαβήτη πριν και μετά την εφαρμογή των μεθόδων διόρθωσης.

Τα αποτελέσματα κατά την εφαρμογή της μεθόδου διόρθωσης FDR ήταν 61 στατιστικά σημαντικά γονίδια για το σακχαρώδη διαβήτη τύπου 1 και 12 στατιστικά σημαντικά γονίδια στο σακχαρώδη διαβήτη τύπου 2. Στο Πίνακα 7 δίνεται ο αριθμός των στατιστικά σημαντικών γονιδίων με επίπεδα σημαντικότητας 1% και 5% και κατόπιν της εφαρμογής των μεθόδων διόρθωσης και για τους δύο τύπους διαβήτη ενώ στους πίνακες 8 και 9 παρουσιάζονται τα στατιστικά σημαντικά γονίδια που προέκυψαν από τη μέθοδο διόρθωσης FDR για κάθε τύπο διαβήτη με το αντίστοιχο p-value.

Κατά τον εντοπισμό των στατιστικά σημαντικών γονιδίων πραγματοποιήθηκε έλεγχος αυτών με τα γονίδια που εμπλέκονται με τη νόσο του σακχαρώδη διαβήτη τα οποία είναι καταχωρημένα στη βάση δεδομένων της [GWAS catalog](#). Για το διαβήτη τύπου 1 εντοπίστηκε ένα κοινό το γονίδιο από αυτά που προέκυψαν από τη μετα-ανάλυση (με επίπεδα σημαντικότητας 5%) με αυτό της βάσης. Συγκεκριμένα το BACH2 ταυτίστηκε με ένα από αυτά που έχουν καταχωρηθεί επίσημα ως γονίδια που εμπλέκονται στην εμφάνιση της νόσου.



Εικόνα 21: Διάγραμμα Venn για τα στατιστικά σημαντικά γονίδια της μετα-ανάλυσης και της βάσης GWAS catalog.

Ωστόσο για το διαβήτη τύπου 2 δεν προέκυψε κάτι ανάλογο για τα στατιστικά σημαντικά γονίδια της μετα-ανάλυσης με αυτά της βάσης.

Πίνακας 7: Τα στατιστικά σημαντικά γονίδια με επίπεδα σημαντικότητας 1% και 5% και κατόπιν της εφαρμογής των μεθόδων διόρθωσης.

	p-value<0.05	p-value<0.01	FDR	bonferroni	sidak	holm	holland
Type 1 Diabetes	1626	531	61	9	9	9	9
Type 2 Diabetes	1447	457	12	2	2	2	2

Πίνακας 8: Τα στατιστικώς σημαντικά γονίδια κατά την εφαρμογή της μεθόδου διόρθωσης FDR(0.01) για το Σακχαρώδη διαβήτη τύπου 1.

genename	p-value	genename	p-value	genename	p-value	genename	p-value	genename	p-value
kctd12	2.7e-06	loc649255	.000039	ctrc	.000131	st20as1	.000213	gem	.000308
loc100287902	5.6e-06	tenc1	.000045	loc645485	.000133	snpc5	.000215	gtf2h3	.000313
sesn2	7.0e-06	sh3yl1	.000046	znf438	.000137	rps6ka3	.000217	loc643824	.000322
dlgap2	.000014	pno1	.000048	c6orf226	.000141	gng7	.00026	ssx3	.000334
bhmt	.00002	dock2	.000052	scap	.000143	sh3gl1p1	.000261	loc100129827	.000337
efs	.00002	llgl2	.000056	loc389036	.000161	loc643293	.000265	anxa2p3	.000338
loc441436	.000024	dpy19l2p4	.000072	psphp1	.000163	loc339751	.000269	nap1l3	.00034
ero1a	.00003	loc641364	.000072	sorbs1	.000166	rhcg	.000278	loc731050	.000357
taf6	.000031	loc644347	.000102	znf709	.000179	dok5	.000286	hsa13	.00037
loc728946	.000034	ndst3	.000104	asah2	.00018	spdef	.000287		
loc100289610	.000036	loc644841	.00012	atp5ep2	.000181	hyd1n2	.000287		
hist1h3c	.000037	stard5	.000123	rbm48	.000184	loc649371	.000297		
loc732450	.000037	tp73as1	.00013	snrnp40	.0002	oxa1l	.000299		

Πίνακας 9: Τα στατιστικώς σημαντικά γονίδια κατά τη μέθοδο διόρθωσης FDR(0.05) για το Σακχαρώδη διαβήτη τύπου 2.

genename	p
arf3	1.5e-06
nlrp8	2.2e-06
cfap47	4.8e-06
oxer1	5.4e-06
krt85	8.4e-06
loc652791	.00001
mlt1	.000015
adam12	.000016
piga	.000019
lpin1	.00002
tdgf1	.000023
mtss1	.000024

Εν συνεχεία παραθέτονται στους Πίνακες 10 και 11, τα 20 γονίδια με το μικρότερο p-value. Τα στατιστικώς σημαντικά γονίδια, κατά σύμβαση, παρουσιάζονται με 1.

Πίνακας 10: Τα 20 γονίδια με το μικρότερο p-value με τα αποτελέσματα των μεθόδων διόρθωσης για τα συγκεκριμένα γονίδια για το σακχαρώδη διαβήτη τύπου 1.

genename	p-value	fdr	bonf	sidak	holm	holland
kctd12	2.7e-06	1	1	1	1	1
loc100287902	5.6e-06	1	1	1	1	1
sesn2	7.0e-06	1	1	1	1	1
dlgap2	.000014	1	1	1	1	1
bhmt	.00002	1	1	1	1	1
efs	.00002	1	1	1	1	1
loc441436	.000024	1	1	1	1	1
ero1a	.00003	1	1	1	1	1
taf6	.000031	1	1	1	1	1
loc728946	.000034	1	0	0	0	0
loc100289610	.000036	1	0	0	0	0
hist1h3c	.000037	1	0	0	0	0
loc732450	.000037	1	0	0	0	0
loc649255	.000039	1	0	0	0	0
tenc1	.000045	1	0	0	0	0
sh3yl1	.000046	1	0	0	0	0
pno1	.000048	1	0	0	0	0
dock2	.000052	1	0	0	0	0
llgl2	.000056	1	0	0	0	0
dpy19l2p4	.000072	1	0	0	0	0

Πίνακας 11: Τα 20 στατιστικώς σημαντικά γονίδια με το μικρότερο p-value με τα αποτελέσματα των μεθόδων διόρθωσης για τα συγκεκριμένα γονίδια για το σακχαρώδη διαβήτη τύπου 2.

genename	p-value	bonf	sidak	holm	holland	fdr(0.05)
arf3	1.5e-06	1	1	1	1	1
nlrp8	2.2e-06	1	1	1	1	1
cfap47	4.8e-06	0	0	0	0	1
oxer1	5.4e-06	0	0	0	0	1
krt85	8.4e-06	0	0	0	0	1
loc652791	.00001	0	0	0	0	1
mllt1	.000015	0	0	0	0	1
adam12	.000016	0	0	0	0	1
piga	.000019	0	0	0	0	1
lpin1	.00002	0	0	0	0	1
tdgf1	.000023	0	0	0	0	1
mtss1	.000024	0	0	0	0	1
linc00189	.000031	0	0	0	0	0
krtap41	.000034	0	0	0	0	0
rfesd	.000036	0	0	0	0	0
EIF5B	.000044	0	0	0	0	0
TMEM204	.000044	0	0	0	0	0
ERCC4	.00006	0	0	0	0	0
SUSD2	.000061	0	0	0	0	0
EPB41L3	.000065	0	0	0	0	0

Τα στατιστικά σημαντικά γονίδια που προέκυψαν από τη μέθοδο FDR τοποθετήθηκαν σαν είσοδοι στις πλατφόρμες bioCompedium, Panther και STRING προκειμένου να ληφθούν πληροφορίες σχετικά με τη μοριακή λειτουργία, τις βιολογικές διεργασίες, τα βιοχημικά μονοπάτια στα οποία εμπλέκονται αλλά την δημιουργία δικτύου των πιθανών αλληλεπιδράσεων μεταξύ τους.

4.4 Τα δεδομένα ως είσοδο στις πλατφόρμες bioCompedium, Panther και STRING:

Τα στατιστικώς σημαντικά γονίδια που προέκυψαν μετά την εφαρμογή των μεθόδων διόρθωσης του p-value, έπρεπε να χαρακτηριστούν ώστε να εξεταστεί περαιτέρω η συσχέτισή τους με την εμφάνιση της νόσου. Έτσι οι λίστες των γονιδίων εισήχθησαν αρχικά στη πλατφόρμα της bioCompedium για να εξετασθούν οι μοριακές λειτουργίες, οι βιολογικές διεργασίες και τα βιοχημικά μονοπάτια στα οποία εμπλέκονται. Για την εξακρίβωση των αποτελεσμάτων της πλατφόρμας της bioCompedium, πραγματοποιήθηκε ανάλυση των στατιστικά σημαντικών γονιδίων και στη πλατφόρμα της Panther. Η πλατφόρμα PANTHER (Protein Analysis Through Evolutionary Relationships), σχεδιάστηκε για να ταξινομεί τις πρωτεΐνες (και τα γονίδια τους), προκειμένου να διευκολυνθεί η ανάλυση υψηλής απόδοσης. Έχει παρόμοια λειτουργία με τη πλατφόρμα bioCompedium.

Η βιολογική βάση δεδομένων STRING παρέχει τη δυνατότητα της δημιουργίας δικτύου αλληλεπιδράσεων που αναπτύσσονται μεταξύ των γονιδίων που εισάγονται και χρησιμοποιήθηκε προκειμένου να οπτικοποιηθούν οι πιθανές σχέσεις των στατιστικά σημαντικών γονιδίων. Οι σχέσεις έχουν διαφορετικούς χρωματισμούς και κάθε χρώμα αναφέρεται σε συγκεκριμένη αλληλεπίδραση.

4.4.1 Για το σακχαρώδη διαβήτη τύπου 1:

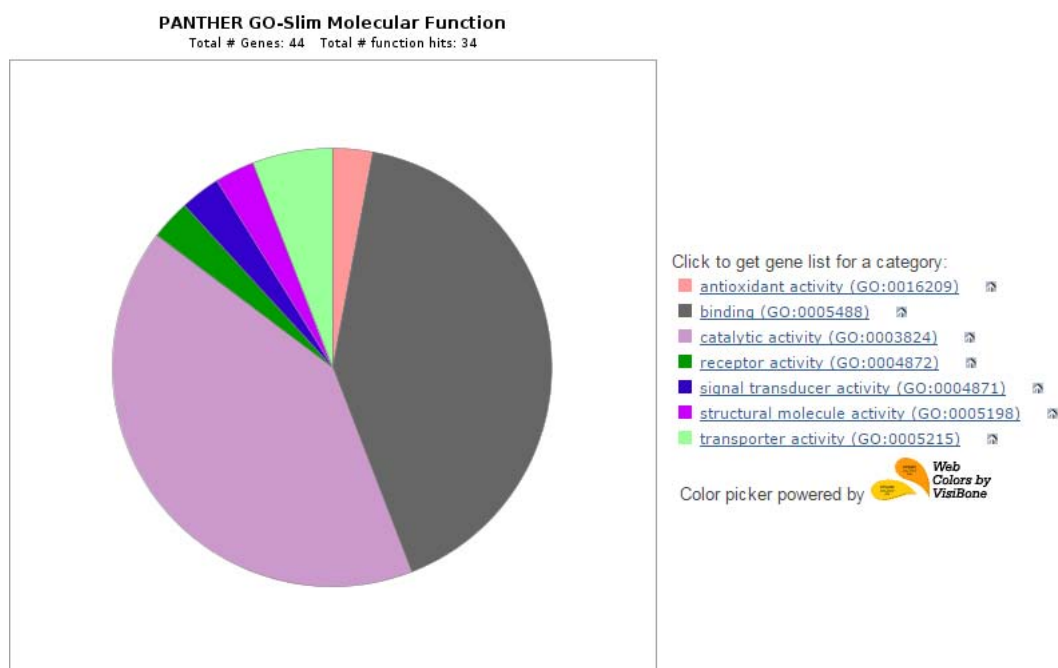
Στη περίπτωση του σακχαρώδη διαβήτη τύπου 1, τα στατιστικά σημαντικά γονίδια κατά τη μέθοδο FDR εισήχθησαν στη πλατφόρμα. Στο σύνολό τους αριθμούσαν 61. Ωστόσο κατά την εισαγωγή, κάποια από αυτά δεν αναγνωρίστηκαν από τις πλατφόρμες Panther και biocompare. Επιπρόσθετα, μετά από έλεγχο στη βάση δεδομένων της Pubmed για τα γονίδια που δεν αναγνωρίστηκαν στις πλατφόρμες, δεν επιστράφηκε κάποια πληροφορία. Αυτό επειδή πολλά από αυτά (τα γονίδια) είτε δεν ήταν γνωστά («uncharacterized») είτε αφορούσαν «γενετικούς τόπους» (παραδείγματος χάριν: loc100129827, loc643824, loc644347 κ.ο.κ). Τελικά από τα αρχικά 61, μόνο για τα 44 από αυτά ληφθήκαν πληροφορίες από τη Panther και 41 από τη biocompare.

Τα γονίδια που δεν αναγνωρίστηκαν από τις 2 πλατφόρμες αλλά και που δεν επιστράφηκαν για αυτά πληροφορίες ούτε από τη βάση δεδομένων της Pubmed, παρουσιάζονται παρακάτω στο Πίνακα 12.

Πίνακας 12: Τα στατιστικά σημαντικά γονίδια του σακχαρώδη διαβήτη τύπου 1 για τα οποία δεν επιστράφηκαν πληροφορίες από τις πλατφόρμες.

Δεν αναγνωρίστηκαν
hist1h3c
tp73as1
loc100129827
st20as1
loc643293
loc644347
loc644841
loc645485
loc649255
loc728946
loc731050
hydin2
loc100287902
loc100289610
loc339751
loc641364
sh3glp1

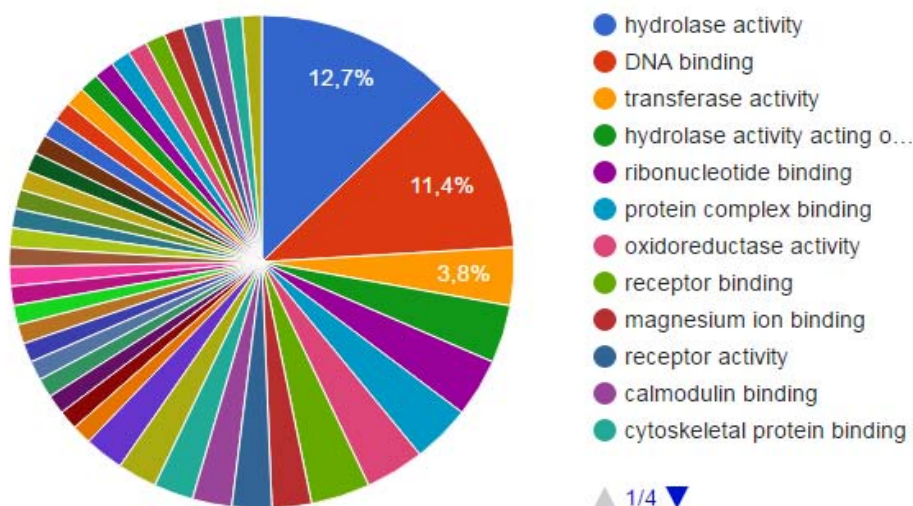
Για τη μοριακή λειτουργία: Τα 44 γονίδια που εισήχθησαν στη Panther εμπλέκονταν σε αρκετές λειτουργίες. Οι πιο επικρατούσες ήταν η λειτουργία της δέσμησης (binding) και η καταλυτική δράση αλλά και άλλες λειτουργίες όπως η αντιοξειδωτική δράση, η μεταφορά κ.α. (Εικόνα 22).



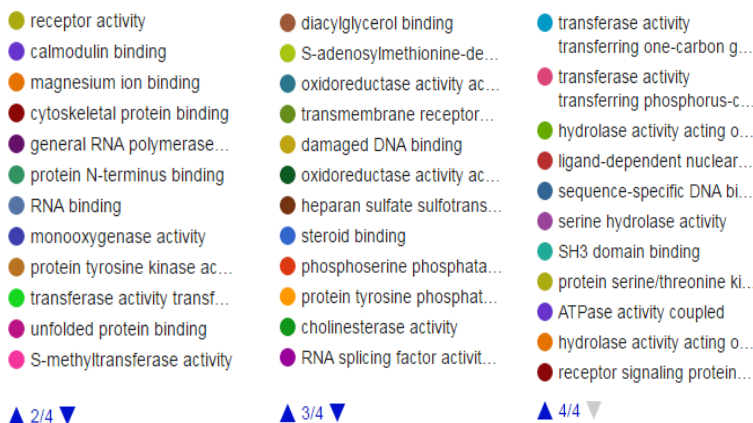
Εικόνα 22: Οι μοριακές λειτουργίες στις οποίες εμπλέκονται τα 44 γονίδια.

Στη πλατφόρμα της biocompedium για τα 41 γονίδια που εισήχθησαν επιστράφηκαν πληροφορίες μόνο για τα 32 από αυτά για τη μοριακή λειτουργία. Στην Εικόνα 23 παρουσιάζεται το pie-chart της μοριακής λειτουργίας των 32 γονιδίων. Και οι δύο πλατφόρμες επέστρεψαν περίπου τις ίδιες λειτουργίες.

Molecular Function Gene Ontology Enrichment Results

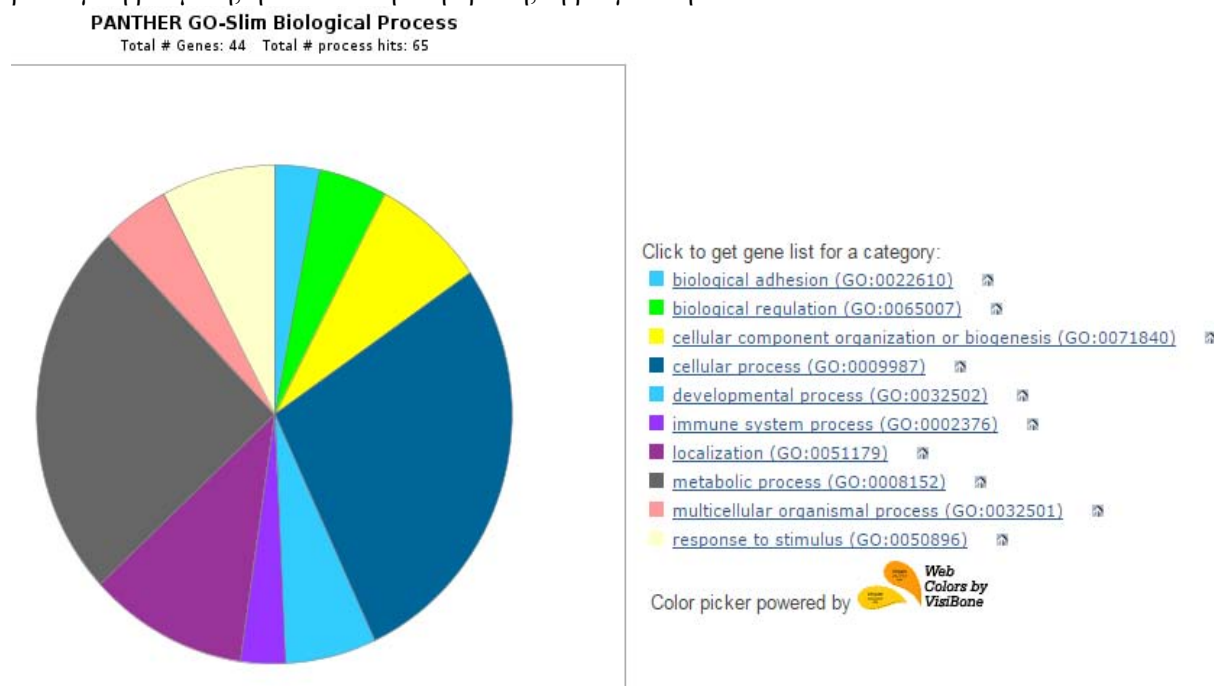


Εικόνα 23: Οι μοριακές λειτουργίες των 32 γονιδίων από τη πλατφόρμα της biocompedium.



Εικόνα 24: Οι υπόλοιπες μοριακές λειτουργίες στις οποίες εμπλέκονται τα 32 γονίδια για το σακχαρώδη διαβήτη τύπου 1.

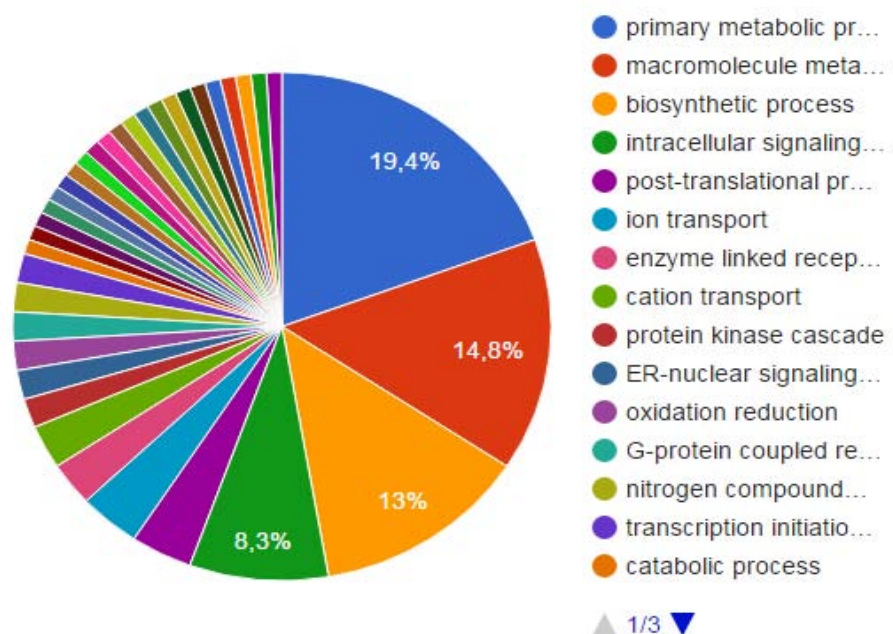
Για τη βιολογική διεργασία: τα αποτελέσματα της Panther για τα 44 γονίδια παρουσιάζονται στην Εικόνα 25. Οι βιολογικές διεργασίες στις οποίες εμπλέκονται είναι η κυτταρική και η μεταβολική διεργασία κατά κόρον αλλά και άλλες διεργασίες όπως είναι η βιολογική ρύθμιση, η ανοσοποιητική δράση, η βιογένεση.



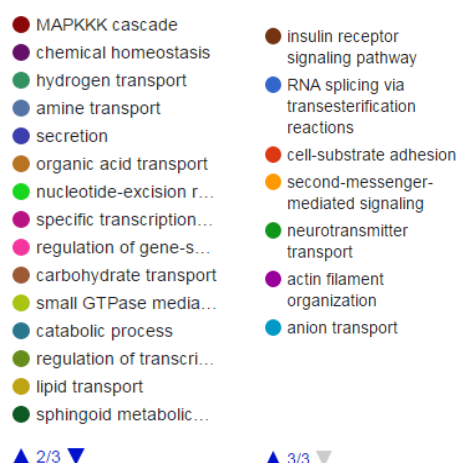
Εικόνα 25: Οι βιολογικές διεργασίες στις οποίες εμπλέκονται τα 44 γονίδια.

Για τη πλατφόρμα της biocompare επιστράφηκαν πληροφορίες για τα 28 γονίδια από τα 41 που αναγνωρίστηκαν. Το μεγαλύτερο ποσοστό συμμετοχής εντοπίστηκε στη πρώιμη μεταβολική διεργασία (21 γονίδια), τη μακρομοριακή μεταβολική διεργασία (16 γονίδια), τη βιοσύνθεση (14 γονίδια) και την ενδοκυττάρια σηματοδότηση αλληλουχίας (9 γονίδια). Τα αποτελέσματα εμφανίζονται στην Εικόνα 26. Το γεγονός της συμμετοχής των γονιδίων σε μεταβολικές διεργασίες οδηγεί στο συμπέρασμα ότι παίζουν σημαντικό ρόλο στο μεταβολισμό και κατ' επέκταση στην πιθανή εμφάνιση μεταβολικών ασθενειών.

Biological Process Gene Ontology Enrichment Results



Εικόνα 26: Οι βιολογικές διεργασίες στις οποίες εμπλέκονται τα 28 γονίδια.



Εικόνα 27: Οι υπόλοιπες βιολογικές διεργασίες στις οποίες εμπλέκονται τα 28 γονίδια για το σακχαρώδη διαβήτη τύπου 1.

Βιοχημικά-Μεταβολικά μονοπάτια (Pathway information):

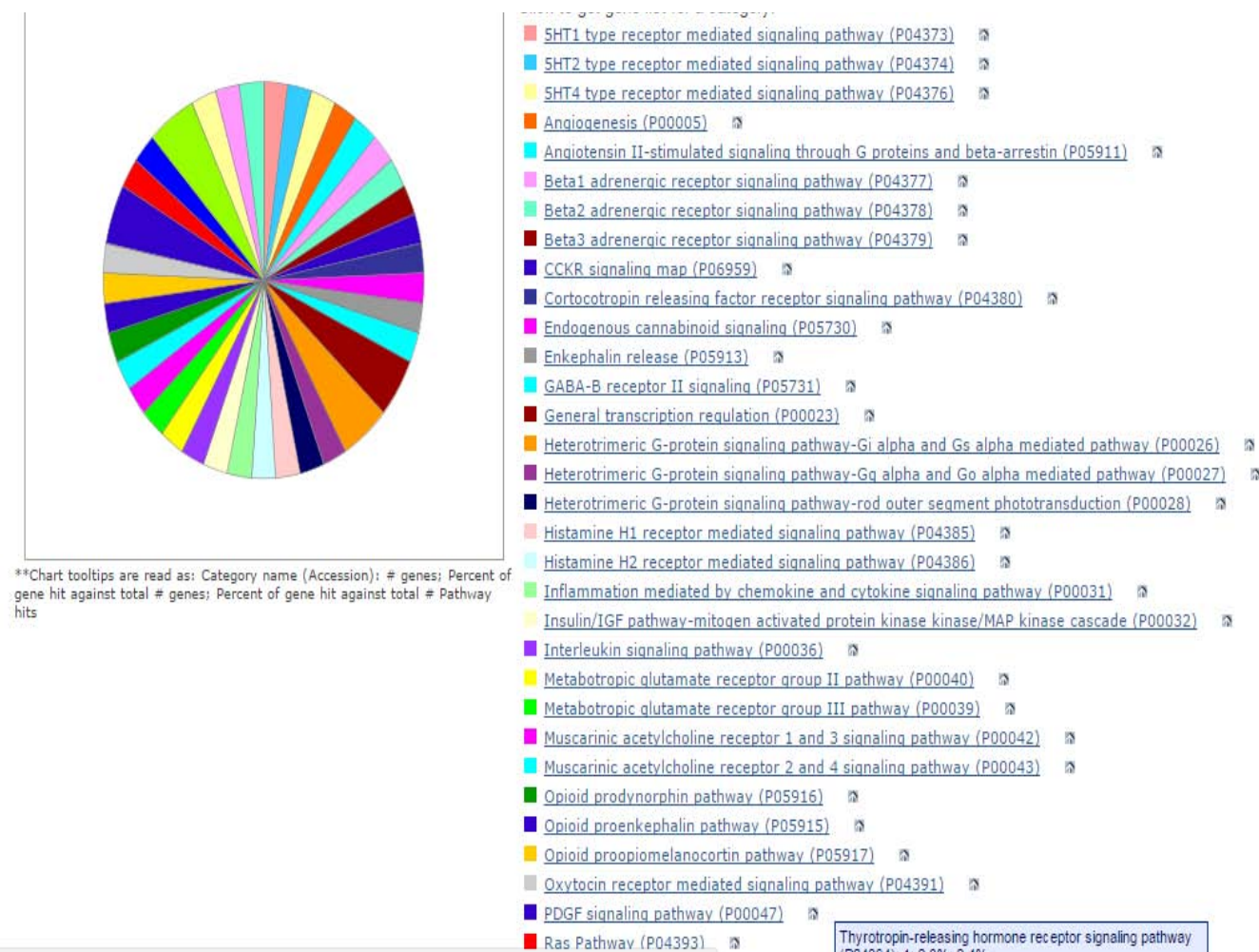
Από το πίνακα που προκύπτει, απεικονίζονται 29 βιοχημικά/μεταβολικά μονοπάτια σύμφωνα με τη βάση δεδομένων της Kegg (μεταβολικό μονοπάτι θεωρείται μια σειρά από χημικές αντιδράσεις μέσα στο κύτταρο σε διαφορετικές χρονικές καταστάσεις). Μερικά από αυτά είναι ο ερυθριματώδης λύκος, οι βασικοί μεταγραφικοί παράγοντες, η νόσος του Parkinson, η ασθένεια του Alzheimer και του Huntington, διάφορες σηματοδοτήσεις, ο μεταβολισμός διαφόρων ουσιών και άλλα (Εικόνα 28).

KEGG Pathway ID	KEGG Pathway Name	Adjusted P-Value	Gene Name	KEGG Gene	Ensembl Gene
hsa05322	Systemic lupus erythematosus	3.0916e-14	HIST1H3A	8350	ENSG00000196532
			HIST1H3D	8351	ENSG00000196532
			HIST1H3C	8352	ENSG00000196532
			HIST1H3E	8353	ENSG00000196532
			HIST1H3I	8354	ENSG00000196532
			HIST1H3G	8355	ENSG00000196532
			HIST1H3J	8356	ENSG00000196532
			HIST1H3H	8357	ENSG00000196532
			HIST1H3B	8358	ENSG00000196532
hsa03022	Basal transcription factors	9.7003e-03	HIST1H3F	8968	ENSG00000196532
			TAF6	6878	ENSG00000106290
hsa00260	Glycine, serine and threonine metabolism	9.7003e-03	GTF2H3	2967	ENSG00000111358
			BHMT	635	ENSG00000145692
hsa00600	Sphingolipid metabolism	9.7003e-03	PSPH	5723	ENSG00000146733
			ASA2H	56624	ENSG00000188611
hsa04062	Chemokine signaling pathway	1.1844e-01	ASA2H2C	653365	ENSG00000188611
			DOCK2	1794	ENSG00000134516
hsa03060	Protein export	1.3366e-01	GNG7	2788	ENSG00000213611
			OXA1L	5018	ENSG00000155463
hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate	1.3711e-01	NDST3	9348	ENSG00000164100
hsa04150	mTOR signaling pathway	1.5484e-01	RPS6KA3	6197	ENSG00000177189
hsa00270	Cysteine and methionine metabolism	1.5484e-01	BHMT	635	ENSG00000145692
hsa03420	Nucleotide excision repair	1.5484e-01	GTF2H3	2967	ENSG00000111358
hsa05110	Vibrio cholerae infection	1.5484e-01	ERO1L	30001	ENSG00000197930
hsa00983	Drug metabolism - other enzymes	1.5484e-01	CES7	221223	ENSG00000159398
hsa04720	Long-term potentiation	1.6048e-01	RPS6KA3	6197	ENSG00000177189
hsa03320	PPAR signaling pathway	1.6048e-01	SORBS1	10580	ENSG00000095637
hsa04914	Progesterone-mediated oocyte maturation	1.6048e-01	RPS6KA3	6197	ENSG00000177189
hsa04520	Adherens junction	1.6048e-01	SORBS1	10580	ENSG00000095637
hsa04115	p53 signaling pathway	1.6048e-01	SESN2	83667	ENSG00000130766
hsa04666	Fc gamma R-mediated phagocytosis	1.7275e-01	DOCK2	1794	ENSG00000134516
hsa04722	Neurotrophin signaling pathway	1.7935e-01	RPS6KA3	6197	ENSG00000177189
hsa03040	Spliceosome	1.7935e-01	WDR57	9410	ENSG00000060688
hsa04910	Insulin signaling pathway	1.7935e-01	SORBS1	10580	ENSG00000095637
hsa05012	Parkinson's disease	1.7935e-01	ATP5E	514	ENSG00000124172
hsa04114	Oocyte meiosis	1.7935e-01	RPS6KA3	6197	ENSG00000177189
hsa04530	Tight junction	1.7935e-01	LLGL2	3993	ENSG00000073350
hsa05010	Alzheimer's disease	2.0392e-01	ATP5E	514	ENSG00000124172
hsa05016	Huntington's disease	2.1657e-01	ATP5E	514	ENSG00000124172
hsa00190	Oxidative phosphorylation	2.4205e-01	ATP5E	514	ENSG00000124172
hsa04010	MAPK signaling pathway	2.6456e-01	RPS6KA3	6197	ENSG00000177189
hsa04080	Neuroactive ligand-receptor interaction	2.6456e-01	HRH3	11255	ENSG00000101180

Εικόνα 28:Τα 29 βιοχημικά μονοπάτια Kegg για τα 41 γονίδια.

Για τα 44 στατιστικώς σημαντικά γονίδια που εισήχθησαν στη πλατφόρμα της Panther προέκυψε ότι εμπλέκονται σε 37 βιοχημικά/μεταβολικά μονοπάτια (Εικόνα 29).

Παρά τη συμμετοχή των γονιδίων σε διάφορες μεταβολικές δραστηριότητες (κάτι που υποδεικνύει τη συσχέτιση αυτών με την πιθανή εμφάνιση της νόσου του διαβήτη) δε βρέθηκε κάποια ανάμειξη αυτών με βιοχημικά/μεταβολικά μονοπάτια της νόσου του σακχαρώδη διαβήτη.

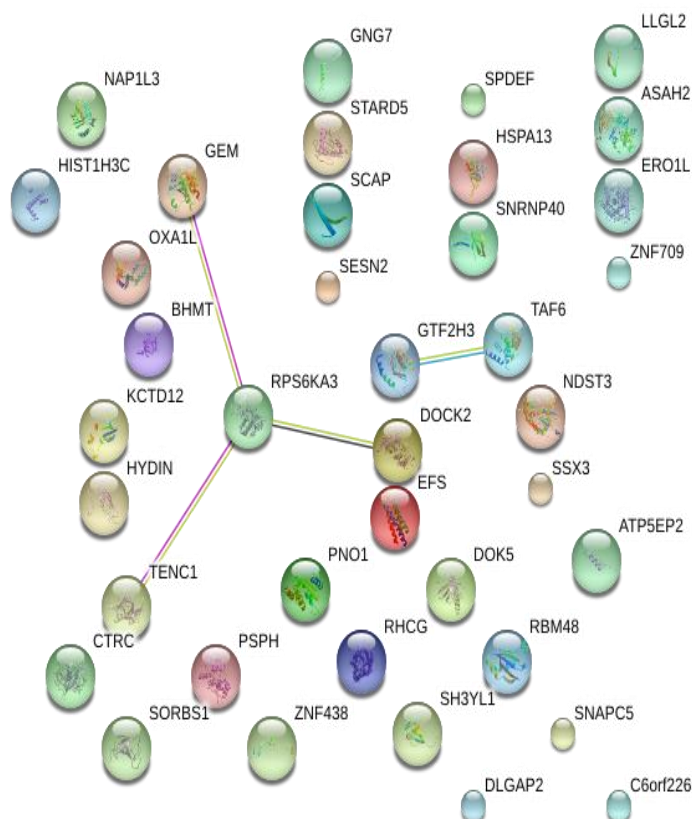


Εικόνα 29: Τα βιοχημικά μονοπάτια των 44 γονιδίων. Δίπλα από το pie chart εμφανίζεται μέρος των μονοπατιών.

Δίκτυο αλληλεπίδρασης σύμφωνα με τη βιολογική βάση STRING:

Για να υπάρχει μία συγκεντρωτική «εικόνα» των γονιδίων που μελετώνται δημιουργήθηκε, με τη χρήση της βάσης STRING, ένα δίκτυο αλληλεπιδράσεων στο οποίο εμφανίζονται όλα τα γονίδια και οι μεταξύ τους συσχετίσεις. Από ότι είναι φανερό τα γονίδια αυτά δεν έχουν κάποια έντονη αλληλεπίδραση μεταξύ τους (Εικόνα 30).

Πιο συγκεκριμένα, τα γονίδια GEM & RPS6KA3 και τα RPS6KA3 & TENC1 φαίνεται να εμφανίζονται μαζί σε πειράματα που έχουν γίνει και να βρίσκονται σε γειτονικούς γονιδιακούς τόπους, τα RPS6KA3 & DOCK 2 φαίνεται να συνεκφράζονται και να βρίσκονται σε γειτονικούς γονιδιακούς τόπους ενώ τα GTF2H3 & TAF6 φαίνεται να βρίσκονται μαζί σε εξειδικευμένες βάσεις και σε γειτονικούς γονιδιακούς τόπους.



Εικόνα 30: Δίκτυο αλληλεπιδράσεων των 39 γονιδίων (από τα 61) που αναγνωρίστηκαν από τη βάση STRING.

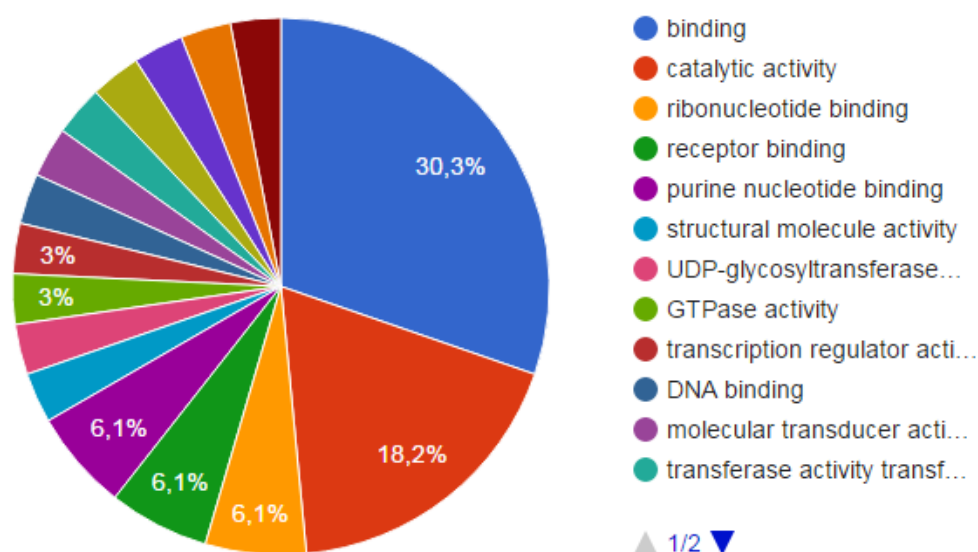
4.4.2 Για το σακχαρώδη διαβήτη τύπου 2:

Για τη μοριακή λειτουργία:

Μετά την εφαρμογή της μεθόδου διόρθωσης του p-value $FDR(p=0.05)$, βρέθηκαν 12 στατιστικώς σημαντικά γονίδια. Για να εξεταστεί η μοριακή λειτουργία αυτών εισήχθησαν στις 2 online πλατφορμες, τη biocompendium και τη Panther.

Εισάγοντας τα 12 στατιστικώς σημαντικά γονίδια στη bioCompendium, ληφθήκαν πληροφορίες για τα 11. Συγκεκριμένα επικράτησαν κυρίως δύο δραστηριότητες (με τα μεγαλύτερα ποσοστά), η λειτουργία της δέσμευσης (περιλαμβάνει πολλές μορφές) και η καταλυτική δραστηριότητα με αρκετά μεγάλο ποσοστό ενώ συμμετείχαν και σε δραστηριότητες μεταφοράς, υποδοχής και μοριακής δόμησης (Εικόνα 31).

Molecular Function Gene Ontology Enrichment Results



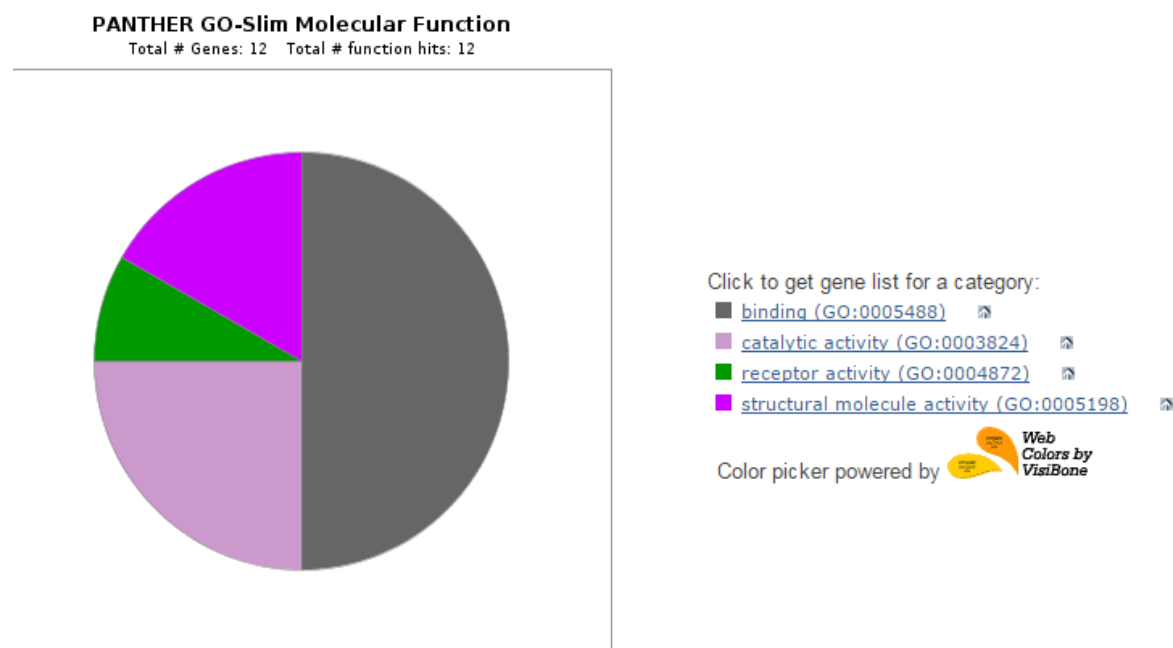
Εικόνα 31: Οι μοριακές λειτουργίες στις οποίες εμπλέκονται τα 11 γονίδια. Δεξιά της εικόνας παρουσιάζεται μέρος των λειτουργιών.

- GTPase activity
- structural molecule activity
- transcription regulator activity
- cytoskeletal protein binding

▲ 2/2 ▼

Εικόνα 32: Οι υπόλοιπες μοριακές λειτουργίες στις οποίες εμπλέκονται τα 11 γονίδια για το σακχαρώδη διαβήτη τύπου 2.

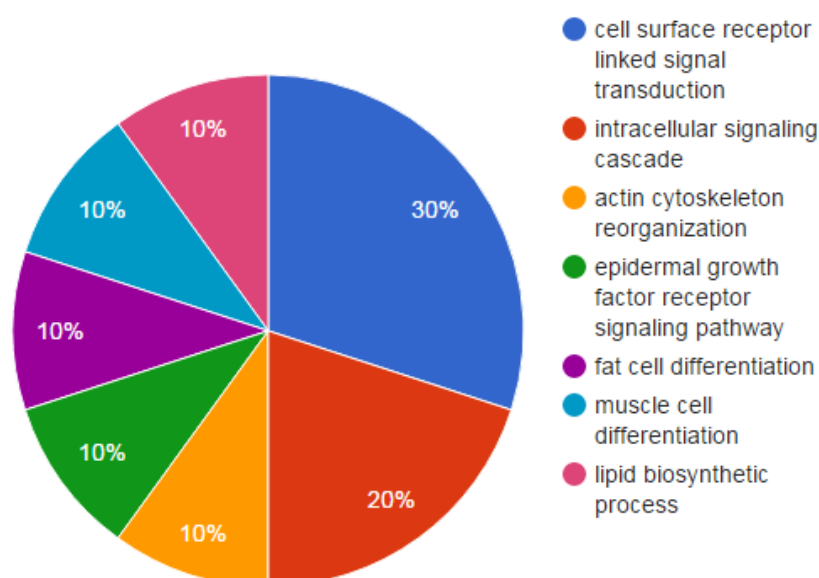
Τα αποτελέσματα της Panther έδειξαν ότι τα 6 γονίδια εμπλέκονται στη δέσμευση, 2 στη μοριακή δόμηση, 3 στη καταλυτική δράση και 1 στη δράση υποδοχέα. Παρακάτω παρουσιάζεται το pie-chart για τη μοριακή λειτουργία των γονιδίων (Εικόνα 33).



Εικόνα 33: Οι μοριακές λειτουργίες στις οποίες εμπλέκονται τα γονίδια.

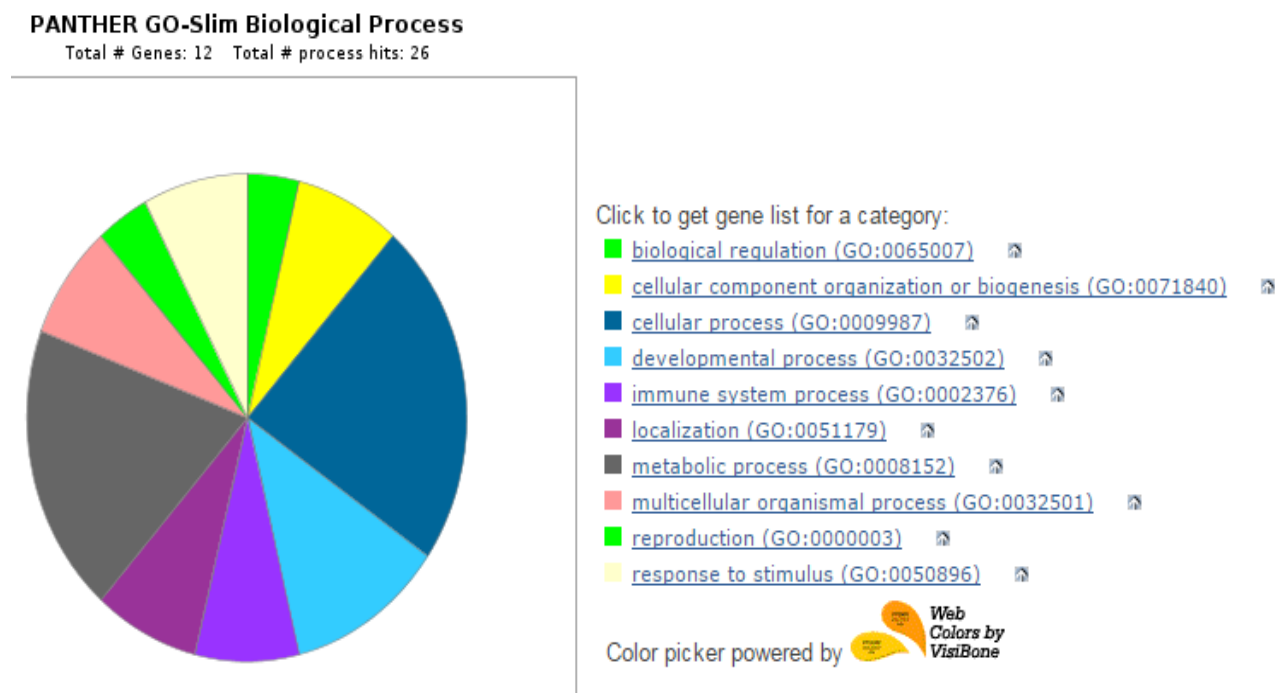
Για τη βιολογική διεργασία: οι πληροφορίες που επιστραφήκαν κατά την είσοδο στη πλατφόρμα της biocompare ήταν για τα 7 από τα 11 γονίδια. Τα αποτελέσματα εμφάνισαν τη συμμετοχή σε 7 διεργασίες. Αναλυτικότερα, 3 στη σηματοδότηση υποδοχέα κυτταρικής μεμβράνης, 2 στην ενδοκυτταρική αλληλουχία σηματοδότησης, 1 στην αναδιοργάνωση του κυτταροσκελετού της ακτίνης, 1 στο μονοπάτι σηματοδότησης του υποδοχέα στον επιδερμικό αυξητικό παράγοντα, 1 στη διαφοροποίηση των λιπιδίων, 1 στη διαφοροποίηση των μυϊκών κυττάρων και 1 στη διεργασία της βιοσύνθεσης των λιπιδίων (Εικόνα 34).

Biological Process Gene Ontology Enrichment Results



Εικόνα 34: Οι βιολογικές διεργασίες στις οποίες εμπλέκονται τα 7 γονίδια.

Κατά την εισαγωγή των 12 γονιδίων στη πλατφόρμα της Panther, τα αποτελέσματα εμφάνισαν τη συμμετοχή τους συνολικά σε 26 βιολογικές διεργασίες με κύρια συμμετοχή στη κυτταρική και μεταβολική διεργασία (Εικόνα 35).



Εικόνα 35: Οι βιολογικές διεργασίες για τα 12 γονίδια.

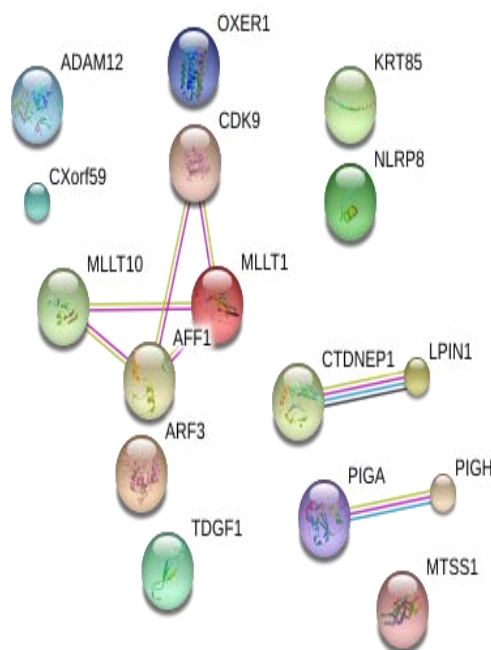
Βιοχημικά/μεταβολικά μονοπάτια (Pathway information):

Στη πλατφόρμα της biocompendium, σύμφωνα με τη βάση δεδομένων της Kegg για τα 11 γονίδια βρέθηκε μόνο ένα βιοχημικό μονοπάτι με συμμετοχή του γονιδίου PIG-A. Πιο συγκεκριμένα το γονίδιο αυτό εμπλέκεται στη βιοσύνθεση των αγκυροβολημένων GPI (glycosylphosphatidylinositol) πρωτεϊνών.

Στη πλατφόρμα της Panther, βρέθηκε η συμμετοχή ενός μόνο γονιδίου σε ένα βιοχημικό/μεταβολικό μονοπάτι, αυτό της νόσου Huntington.

Δίκτυο αλληλεπίδρασης σύμφωνα με τη βάση STRING:

Το δίκτυο αλληλεπιδράσεων, που δημιουργήθηκε από τη βιολογική βάση STRING, των 11 στατιστικώς σημαντικών γονιδίων παρουσιάζεται παρακάτω στην Εικόνα 36.



Εικόνα 36: Δίκτυο αλληλεπιδράσεων μεταξύ των γονιδίων.

Στο συγκεκριμένο δίκτυο που προέκυψε για τα 12 στατιστικά γονίδια δεν εντοπίστηκε κάποια έντονη αλληλεπίδραση μεταξύ τους ούτε η ύπαρξη κλειστών συνόλων. Τα γονίδια MLLT10, MLLT1, CDK9 & AFF1 αλληλεπιδρούν ως γονίδια που εμφανίζονται μαζί σε πειράματα και προέρχονται από textmining, πιθανόν κάποια κοινή προέλευση. Τα CTDNEP1 και LPIN1 έχουν τεσσάρων ειδών αλληλεπιδράσεων, εμφανίζονται μαζί σε πειράματα, προέρχονται από textmining, εντοπίζονται μαζί σε εξειδικευμένες βάσεις και συνεκφράζονται. Τέλος, τα PIGA και PIGH παρουσιάζουν τις ίδιες αλληλεπιδράσεις με τα προηγούμενα με μόνη διαφορά ότι δεν συνεκφράζονται.

4.5 Σύνοψη:

Στη παρούσα διπλωματική εργασία, πραγματοποιήθηκε εκτενής αναζήτηση στη βάση δεδομένων της GEO Datasets προκειμένου να ανακτηθούν τα γονίδια τα οποία παρουσιάζουν κάποια συσχέτιση με την εμφάνιση της νόσου του Σακχαρώδη Διαβήτη. Μετά από ενδελεχή έλεγχο και προσαρμόζοντας τα δεδομένα πραγματοποιήθηκε, με τη βοήθεια του στατιστικού προγράμματος STATA13, ο έλεγχος t-test και bootstrap για κάθε μελέτη ξεχωριστά. Τα αποτελέσματα που προέκυψαν συνενώθηκαν, ξεχωριστά για κάθε τύπο διαβήτη, και πραγματοποιήθηκε η μετ-ανάλυση.

Για την εξαγωγή συμπερασμάτων, αν τελικά εμπλέκεται ή όχι κάποιο από όλα τα γονίδια που μετ-αναλύθηκαν στην εμφάνιση της νόσου, ελέγχθηκε η τιμή του p-value.

Καθώς ο αριθμός των γονιδίων ήταν αρκετά μεγάλος, το επίπεδο σημαντικότητας 5% ή 1% δεν χρησιμοποιήθηκε για να προσδιοριστούν τα στατιστικώς σημαντικά γονίδια αλλά εφαρμόστηκαν μέθοδοι διόρθωσης του p-value.

Για τα στατιστικώς σημαντικά γονίδια που προέκυψαν ότι έχουν συσχέτιση με τη νόσο, διαπιστώθηκε ότι εμπλέκονται σε διάφορους μηχανισμούς του κυττάρου, κυρίως σε μεταβολικές διεργασίες και στη δέσμευση, ωστόσο δεν είναι ξεκάθαρο αν πράγματι σχετίζονται με την εμφάνιση της νόσου. Για να επιτευχθεί ή να ενισχυθεί η απόδειξη της ύπαρξης γονιδίων που σχετίζονται με τη νόσο, θα πρέπει να πραγματοποιηθούν επιπλέον έρευνες και μελέτες.

Βιβλιογραφία:

- (WHO), W. H. O. (2016). "Diabetes country profiles 2016."
- (WHO), W. H. O. (2016). "Global report on diabetes." from http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1, .
- Demeter et al. (2007). "The Stanford Microarray Database: implementation of new analysis tools and open source release of software." Nucleic Acids Res.
- Alexander Sturn, J. Q. a. Z. T. (2002). "Genesis: cluster analysis of microarray data." Oxfords journals **18**.
- B. Samuel Lattimore, S. v. D., M. James C. Crabbe (2005). "GeneMCL in microarray analysis." Computational Biology and Chemistry **29**(5).
- Barrett T, Edgar R. (2006). "Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*." Methods Mol Biol.
- bioCompendium. "The high-throughput experimental data analysis platform." from <http://biocompendium.embl.de/cgi-bin/biocompendium.cgi>.
- Brazma A, P. H., Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res.
- Fangxin Hong and Breitling, F. H. a. R. (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments." Bionformatics, Oxfords Journal.
- Christian Heichinger, C. J. P., Jürg Bähler, Paul Nurse (2006). "Genome-wide characterization of fission yeast DNA replication origins." EMBO Journal.
- Copenhaver, B. S. H. a. M. D. (1987). "An Improved Sequentially Rejective Bonferroni Test Procedure." Biometrics **43**: 7.
- Stekel D. (2003). "Microarray Bioinformatics." Cambridge University Press, UK.
- Dalton L, B. V., Brun M. (2009). "Clustering algorithms: on learning, validation, performance, and applications to genomics." Curr Genomics.
- Efron, T. J. D. a. B. (1996). "Bootstrap Confidence Intervals." Statistical Science **11**: 24.
- Egger, D. S. e. a. (1997). "Bias in meta-analysis detected by a simple, graphical test." BMJ.
- Eric P. Xing , M. I. J., Richard M. Karp (2001). "Feature selection for high-dimensional genomic microarray data." In Proceedings of the Eighteenth International Conference on Machine Learning: 8.
- Eye, A. V. (2003). "Configural Frequency Analysis: Methods, Models, and Applications: Psychology Press."
- Computational Genetics Groups, "Statistical Methods for analysis and meta-analysis of microarrays." Computational Genetics Groups (<http://www.compgen.org>).
- Hochberg, Y. B. a. Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society: 12.
- Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure." Scandinavian Journal of Statistics **6**.
- Hongfang Liu, I. B., and Xin Li (2007). "Microarray probes and probe sets." Front Biosci (Elite Ed).
- Hoheisel JD, (2006). "Microarray technology: beyond transcript profiling and genotype analysis." Nature Reviews Genetics.
- Kerr, K. F. (2007). "Extended analysis of benchmark datasets for Agilent two-color microarrays." BMC Bioinformatics.
- Lee, H. e. a. (2004). "Coexpression analysis of human genes across many microarray data sets." Genome Res.

- Leung YF, C. D. (2003). "Fundamentals of cDNA microarray data analysis." Trends in genetics.
- N. M. Luscombe, D. G., M. Gerstein. (2001). "What is bioinformatics? A proposed definition and overview of the field." Method Inform Med from <https://www.ebi.ac.uk/luscombe/docs/mim-review.pdf>.
- Normand, S. L. (1999). "Meta-analysis: formulating, evaluating, combining, and reporting." STATISTICS IN MEDICINE.
- Panther "The PANTHER (Protein Analysis Through Evolutionary Relationships) Classification System."
- Quackenbush, J. (2001). "Computational genetics: Computational analysis of microarray data." Nature Journal.
- Schena, M. (2003). "Microarray Analysis." John Wiley & Sons, Inc., Hoboken: 664.
- Šidák, Z. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." Journal of the American Statistical Association **62**.
- Sturn A, Q. J., Trajanoski Z., (2002). "Genesis: cluster analysis of microarray data." Bioinformatics.
- Thakkinstian A, M. P., D'Este C, Duffy D, Attia J. (2005). "A method for meta-analysis of molecular association studies." Stat Med.
- Tseng, G. C. (2012). "Comprehensive literature review and statistical considerations for microarray meta-analysis." Nucleic Acids Res.
- Willard M. Freeman, D. J. and R. a. K. E. Vrana (2000). "Fundamentals of DNA Hybridization Arrays for Gene Expression Analysis." BioTechniques **29**.
- Yee Hwa Yang, S. D., Percy Luu, David M. Lin, Vivian Peng, John Ngai and Terence P. Speed (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." Oxford journals **30**.
- Efron B. (1993). "An introduction to the Bootstrap."
- "GWAS Catalog.", <https://www.ebi.ac.uk/gwas/>
- "HGNC HUGO." , <http://www.genenames.org/>
- Daniel R. Rhodes, T. R. B., Mark A. Rubin, Debashis Ghosh, and Arul M. Chinnaiyan (2002). "Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer." CANCER RESEARCH **62**.
- Jonathan A. Ewald, T. M. D., Jeremy P. Cetnar, William A. Ricke (2013). "Expression Microarray Meta-Analysis Identifies Genes Associated with Ras/MAPK and Related Pathways in Progression of Muscle-Invasive Bladder Transition Cell Carcinoma." PLOS.
- Anders, M., Fehlfker, M., Wang, Q., Wissmann, C., Pilarsky, C. Kemmner, W., Höcker, M. (2013). "Microarray meta-analysis defines global angiogenesis-related gene expression signatures in human carcinomas." Molecular Carcinogenesis **52(1)**: 29-38.
- Burguillo, F. J., Martin, J., Barrera, I., Bardsley, W.G. (2010). "Meta-analysis of microarray data: The case of imatinib resistance in chronic myelogenous leukemia." Computational Biology and Chemistry **34(3)**: 184-192.

Παράρτημα 1:

Οι μελέτες που απορρίφθηκαν :

GSE67297	GSE33070	GSE40234	GSE27949	GSE65561
GSE41767	GSE19649	GSE77350	GSE72492	GSE56606
GSE44639	GSE38396	GSE19637	GSE70752	GSE52314
GSE51924	GSE1009	GSE26887	GSE64998	GSE65057
GSE23343	GSE15653	GSE61142	GSE47385	GSE13760
GSE4704	GSE12643	GSE53257	GSE62761	GSE71416
GSE16415	GSE61166	GSE20966	GSE69595	GSE50397
GSE53454	GSE44035	GSE38642	GSE21232	GSE25724
GSE66413	GSE37025	GSE68049	GSE55100	GSE55099
GSE52724	GSE51058	GSE40496	GSE14368	GSE22255
GSE24147	GSE11907	GSE11908	GSE11907	GSE74629
GSE30211	GSE30209	GSE30208	GSE43488	
GSE78891	GSE67775	GSE67774	GSE67773	GSE55311
GSE76161	GSE76899	GSE76268	GSE76189	GSE76065
GSE75685	GSE75669	GSE75678	GSE69600	GSE69889
GSE70961	GSE73418	GSE71730	GSE63423	GSE21891
GSE69658	GSE71099	GSE71102	GSE70528	GSE59421
GSE67543	GSE69421	GSE66785	GSE67279	GSE48278
GSE68571	GSE60760	GSE68226	GSE62117	GSE63887
GSE37084	GSE67567	GSE62219	GSE66360	GSE55645
GSE66175	GSE62372	GSE62370	GSE62832	GSE60424
GSE63981	GSE45856	GSE38267	GSE56081	GSE57896
GSE55465	GSE55464	GSE62523	GSE62500	GSE62499
GSE61769	GSE61714	GSE61129	GSE56781	
GSE37794	GSE30161	GSE35851	GSE30802	GSE35186
GSE35191	GSE38447	GSE32575	GSE34223	GSE37824
GSE37901	GSE28384	GSE37639	GSE30159	GSE32512
GSE26244	GSE32874	GSE32691	GSE35296	GSE31901
GSE35411	GSE32544	GSE19943	GSE32357	GSE24818
GSE28022	GSE28024	GSE27507	GSE27175	GSE27317
GSE21815	GSE31056	GSE23506	GSE30803	GSE30732
GSE30310	GSE28059	GSE27951	GSE29718	GSE24326
GSE29190	GSE29084	GSE25862	GSE24193	GSE26744
GSE21980	GSE26168	GSE26167	GSE26073	GSE25826
GSE24422	GSE24685	GSE25249	GSE23784	GSE16804
GSE14503	GSE24215	GSE23858	GSE18821	
GSE41744	GSE43950	GSE36397	GSE36403	GSE36402
GSE36084	GSE29908	GSE33032	GSE27645	GSE23561

GSE4901	GSE1322	GSE70453	GSE70494	GSE70493
GSE2956	GSE46899	GSE46900	GSE46897	GSE15790
GSE3118	GSE67738	GSE50866	GSE73034	GSE74782
GSE68224	GSE55650	GSE38835	GSE38291	GSE29221
GSE25462	GSE19420	GSE22309	GSE21340	GSE18732
GSE634	GSE121	GSE59363	GSE18470	GSE29226
GSE24290	GSE51546	GSE49524	GSE65737	GSE29231
GSE73408	GSE69528	GSE62003	G98SE341	GSE44314
GSE44313	GSE19790	GSE20553	GSE20067	GSE68526
GSE35713	GSE35712	GSE35711	GSE35725	
GSE56685	GSE60803	GSE52376	GSE55567	GSE55566
GSE42902	GSE30575	GSE58634	GSE44558	GSE57928
GSE57880	GSE57484	GSE50005	GSE48101	GSE29536
GSE54279	GSE46097	GSE48354	GSE48353	GSE53949
GSE29623	GSE29622	GSE29621	GSE28038	GSE40878
GSE36233	GSE40360	GSE52422	GSE51311	GSE51310
GSE50800	GSE43580	GSE42432	GSE50892	GSE35279
GSE50386	GSE42715	GSE39825	GSE49566	GSE47874
GSE47720	GSE40498	GSE42507	GSE45986	GSE45792
GSE45777	GSE43752	GSE43751	GSE43750	GSE32909
GSE34512	GSE42487	GSE42229	GSE42228	GSE42227
GSE42148	GSE42094	GSE42093	GSE35716	
GSE17710	GSE12385	GSE12384	GSE21785	GSE21989
GSE17941	GSE19519	GSE17727	GSE20247	GSE19769
GSE18927	GSE16256	GSE13840	GSE15072	GSE12959
GSE18212	GSE8908	GSE17556	GSE15543	GSE17060
GSE17058	GSE16025	GSE13015	GSE5903	GSE13736
GSE13465	GSE13920	GSE10334	GSE13290	GSE8157
GSE10540	GSE9588	GSE9984	GSE9939	GSE6751
GSE6599	GSE6798	GSE9157	GSE9017	GSE9105
GSE7818	GSE7146	GSE6862	GSE6573	GSE5090
GSE3881	GSE3308	GSE3447	GSE3307	GSE2138
GSE34526	GSE68185	GSE68184	GSE68183	GSE21321
GSE52233	GSE49885	GSE44093	GSE48318	

Παράρτημα 2:

Ο κώδικας STATA που χρησιμοποιήθηκε:

Ο έλεγχος t-test & bootstrap υλοποιήθηκε για κάθε μελέτη χωριστά.

```
set more off

file open meta using results1.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 a5-zzz3 {

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(100) 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
```

Κατόπιν τα αποτελέσματα από κάθε μελέτη ενώθηκαν για κάθε τύπο διαβήτη.

append using "filename.dta"

όπου filename.dta αρχείο που θα προσθέτεται στο προηγούμενο αρχείο (που θα είναι ήδη ανοιχτό). Αφού λοιπόν ενώσαμε τα αρχεία μας για κάθε τύπο διαβήτη, εν συνεχεία εκτελέσαμε το κώδικα της μετα-ανάλυσης.

Ο κώδικας της μετα-ανάλυσης δίνεται ως εξής:

```
encode gene, gen(ngene)
```

```
qui sum ngene
```

```
qui local k=r(max)
```

```
file open metan using finalresults2.txt, write append
```

```
file write metan "gene"
```

```
file write metan " "
```

```
file write metan "ES"
```

```
file write metan " "
```

```
file write metan "se(ES)"
```

```
file write metan " "
```

```
file write metan "z"
```

```
file write metan " "
```

```
file write metan "p"
```

```
file write metan " "
```

```
file write metan "df" _n
```

```
forvalues i=1/'k'{
```

```
metan t rse if ngene==`i',nograph randomi
```

```
file write metan "`i'"
```

```
file write metan " "
```

```
file write metan "`r(ES)'"
```

```
file write metan " "
```

```
file write metan "`r(seES)'"
```

```
file write metan " "
```

```
file write metan "`r(z)'"
```

```
file write metan " "
```

```
file write metan "`r(p_z)'"
```



```
file write metan " "  
  
file write metan ""r(df)"" _n  
  
}  
  
file close metan
```

Αυτό που προκύπτει είναι ένα αρχείο text με τίτλο finalresults(1ή2).txt όπου περιέχει το p-value, standard error && z καθώς και τα γονίδια (με αύξοντα αριθμό). Για την εύρεση των στατιστικά σημαντικών γονιδίων πραγματοποιήθηκε αρχικά:

```
count if p<0.05  
count if p<0.01
```

και στη συνέχεια υλοποιήθηκαν οι μέθοδοι διόρθωσης του p-value :

```
multproc, puncor(0.05) pval(p) meth(simes) rej(fdr)  
  
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)
```