Πανεπιστήμιο Θεσσαλίας
Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική

Ανάλυση Δεδομένων Γονιδιακής Έκφρασης και Πολύπλοκων Βιολογικών Δικτύων

Διδακτορική Διατριβή

Παναγιώτα Κοντού

Λαμία
Δεκέμβριος 2016

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Ανάλυση Δεδομένων Γονιδιακής Έκφρασης και Πολύπλοκων Βιολογικών Δικτύων

Διδακτορική Διατριβή

Παναγιώτα Κοντού

Λαμία, Νοέμβριος 2016

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Στην οικογένειά μου

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

When you feel like quitting, think about why you started

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής, Πανεπιστήμιο Θεσσαλίας, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική

Μαρία Αδάμ
Επίκουρος Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική

Βασίλειος Πλαγιανάκος
Αναπληρωτής Καθηγητής, Πανεπιστήμιο Θεσσαλίας, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική

Άρτεμις Χατζηγεωργίου
Καθηγήτρια**,** Πανεπιστήμιο Θεσσαλίας, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Γεώργιος Σταμούλης
Καθηγητής, Πανεπιστήμιο Θεσσαλίας, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Κωνσταντίνος Παπαλουκάς
Αναπληρωτής Καθηγητής, Πανεπιστήμιο Ιωαννίνων, Τμήμα Βιολογικών Εφαρμογών και Τεχνολογιών

Βασιλική Οικονομίδου
Επίκουρος Καθηγήτρια, ΕΚΠΑ, Τμήμα Βιολογίας

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Ευχαριστίες

Η παρούσα διδακτορική διατριβή εκπονήθηκε από τον Ιούλιο 2011 έως το Νοέμβριο 2016 στο Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας.

Πρώτα από όλους θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Παντελή Μπάγκο, που δέχτηκε να συνεχιστεί η συνεργασία μας μετά το πέρας της πτυχιακής μου εργασίας το 2009 και της μεταπτυχιακής διπλωματικής εργασίας το 2011. Η συνεργασία με τον κο Μπάγκο μου έδωσε εφόδια ανεκτίμητης αξίας για τη συνέχεια της επαγγελματικής μου σταδιοδρομίας και είμαι ιδιαίτερα ευγνώμων στο πρόσωπο του μιας και με έκανε να αγαπήσω τη Βιοπληροφορική και την έρευνα και μου έδωσε το έναυσμα για να ασχοληθώ περαιτέρω με την επιστήμη. Δεν θα ξεχάσω τη φράση που μου έλεγε συνέχεια (και τσακωνόμαστε για το λόγο αυτό…) όταν ήθελα λίγες μέρες ξεκούρασης: «Η έρευνα δεν έχει διακοπές». Τον ευχαριστώ για τη συνεχή καθοδήγηση και την πολύτιμη στήριξη που μου παρείχε όλα αυτά τα χρόνια.

Οφείλω ένα μεγάλο ευχαριστώ στην Επίκουρο Καθηγήτρια του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Μαρία Αδάμ, για την υποστήριξή της, τις πολύτιμες συμβουλές και την προθυμία της κάθε στιγμή να με βοηθήσει τόσο στην εκπόνηση της διατριβής όσο και σε όποια άλλη δυσκολία συνάντησα στη σχολή αλλά και στη ζωή μου γενικότερα. Η κα Αδάμ σαν δεύτερη μαμά ήταν πάντα εκεί για έμενα.

Ακόμη, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή και Πρόεδρο του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική Βασίλειο Πλαγιανάκο και γιατί δέχτηκε να συνεπιβλέψει τη διατριβή μου και για την άριστη συνεργασία που είχαμε όλα αυτά τα χρόνια.

Ευχαριστώ πολύ τα μέλη της εξεταστικής επιτροπής, την Καθηγήτρια κα Αρτέμιδα Χατζηγεωργίου, τον Καθηγητή κ. Γεώργιο Σταμούλη, την Επίκουρο Καθηγήτρια κα Βασιλική Οικονομίδου και τον Αναπληρωτή Καθηγητή κ. Κωνσταντίνο Παπαλουκά, για τις υποδείξεις τους και τον πολύτιμο χρόνο που αφιέρωσαν για τη διόρθωση και την υποστήριξη της παρούσας διατριβής.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω σε όλα τα μέλη του Εργαστηρίου Μοριακής και Υπολογιστικής Βιολογίας και Γενετικής του Πανεπιστημίου Θεσσαλίας, Γεωργία Μπράλιου, Νίκη Δήμου, Κατερίνα Πανταβού και Αθανασία Παυλοπούλου για την πολύτιμη βοήθειά τους και για τη στήριξή τους όποτε τις χρειάστηκα. Αθανασία, σε ευχαριστώ πολύ που με βοήθησες να ξεπεράσω το φόβο μου με τα αγγλικά και το γράψιμο των δημοσιεύσεων και ήσουν πάντα εκεί για εμένα, να συζητήσουμε και να βρούμε λύση σε όποιο πρόβλημα συναντούσα κατά τη διάρκεια της διατριβής. Γκρέτα, σε ευχαριστώ για τα «μυστικά» του εργαστηρίου και των πειραματικών τεχνικών που μου έμαθες και για την προθυμία σου να με βοηθήσεις σε κάθε μου δυσκολία. Σ’ ευχαριστώ για την εμπιστοσύνη και τη στήριξή σου. Το βραδινό μας περπάτημα μετά τη δύσκολη μέρα στο εργαστήριο ήταν το καλύτερο δώρο που θα μπορούσες να μου κάνεις.

Επίσης, ευχαριστώ πολύ τη φίλη μου Ευφροσύνη–Άλκηστη Παρασκευοπούλου-Κόλλια για τη στήριξή της, ιδιαίτερα την τελευταία χρονιά που ήταν η πιο δύσκολη για μένα. Χωρίς αυτήν η εκπλήρωση του διδακτορικού θα ήταν ακόμη ένα μακρόπνοος στόχος. Κάθε φορά που σκεφτόμουν να τα παρατήσω ήταν εκεί να μου υπενθυμίζει τον αρχικό μου στόχο. Στη συνέχεια, θα ήθελα να ευχαριστήσω τη Δρ. και φίλη μου, πάνω απ’ όλα, Vicky G. Tsirkone για την εμπιστοσύνη, τη βοήθεια και για την ψυχολογική υποστήριξη όλα αυτά τα χρόνια. Vicky, για εμένα είσαι πρότυπο…

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδερφό μου για την αγάπη τους, την υπομονή τους και τη συμπαράσταση τους όλα τα χρόνια των σπουδών μου. Μαμά, Μπαμπά, σας ευχαριστώ πολύ που ήσασταν πάντα δίπλα μου, που με στηρίζατε τόσο με λόγια όσο και με πράξεις. Χωρίς εσάς δεν θα ήμουν ο άνθρωπος που είμαι σήμερα.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Τέλος, θα ήθελα να αναφέρω ότι κατά τη διάρκεια της εκπόνησης της παρούσας διδακτορικής διατριβής συμμετείχα και στα ερευνητικά προγράμματα: 1) *«Ενοποίηση δεδομένων από διαφορετικές πηγές: μια σύνθεση της επιδημιολογίας με τη βιοπληροφορική, με εφαρμογές στις πολυπαραγοντικές ασθένειες»* το οποίο συγχρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους στο πλαίσιο της πράξης "ΑΡΙΣΤΕΙΑ ΙΙ" του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Διά Βίου Μάθηση» και 2) *«Ολοκληρωμένη μελέτη του τρόπου δράσης μεμβρανικών υποδοχέων και της εμπλοκής τους σε ασθένειες με σύγχρονες μεθόδους βιοπληροφορικής»*, το οποίο συγχρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης-ΕΤΠΑ) και από εθνικούς πόρους στο πλαίσιο της δράσης εθνικής εμβέλειας «ΣΥΝΕΡΓΑΣΙΑ». Σε αυτό το σημείο καθίσταται σημαντικό να ευχαριστήσω όλους όσους συνέδραμαν στο να υλοποιηθούν τα άνωθεν προγράμματα και με τους οποίους συνεργάστηκα αποτελεσματικά και επιτυχώς. Σε αυτούς περιλαμβάνονται τόσο τα μέλη της ερευνητικής ομάδας του Εργαστηρίου Μοριακής και Υπολογιστικής Βιολογίας και Γενετικής του Πανεπιστημίου Θεσσαλίας τα οποία ανέφερα ήδη, όσο και ο Ομότιμος Καθηγητής κ. Σταύρος Χαμόδρακας και η ερευνητική του ομάδα στο ΕΚΠΑ με τους οποίους συνεργάστηκα στο δεύτερο από τα παραπάνω προγράμματα.

Παναγιώτα Κοντού
Λαμία, Δεκέμβριος 2016

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Δημοσιεύσεις σε επιστημονικά περιοδικά

1. Pantavou K, Braliou GG, Kontou PI, Dimou NL, Bagos PG. **A meta-analysis of FZD3 polymorphisms and their association with schizophrenia.** *Psychiatric genetics*, 26 (2016) 272-280

2. Braliou GG, Grigoriadou AM, Kontou PI, Bagos PG. **The role of genetic polymorphisms of the Renin-Angiotensin System in renal diseases: A meta-analysis.** *Comput Struct Biotechnol J.* 2014 Jun 11;10(16):1-7

3. Braliou GG, Pantavou KG, Kontou PI, Bagos PG**. Polymorphisms of the CD24 Gene Are Associated with Risk of Multiple Sclerosis: A Meta-Analysis.** *International Journal of Molecular Sciences*. 2015

4. Kontou PI, Pavlopoulou A, Braliou GG, Bagos PG (2016). **Identification of differentially expressed genes in myocardial infarction: a Meta-analysis** (submitted)

5. Kontou PI, Pavlopoulou A, Dimou NL, Pavlopoulos GA, Bagos PG. **Network analysis of genes and their association with diseases.** *Gene*, 2016, 590(1), 68-78.

6. Kontou PI, Pavlopoulou A, Dimou NL, Pavlopoulos GA, Bagos PG **Data and programs in support of network analysis of genes and their association with diseases.** *Data in Brief*, 8 (2016) 1036-1039

7. Kontou PI, Pavlopoulou A, Bagos PG. (2016). **Methods of Analysis and Meta-Analysis for Identifying Differentially Expressed Genes.** *Methods Mol Biol* (in press)

8. Kontou PI, Pavlopoulou A, Dimou NL, Theodoropoulou M, Braliou GG, Tsaousis G, Pavlopoulos GA, Hamodrakas SI, Bagos PG. **The human GPCR signal transduction network.** (submitted)

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Περίληψη

Η παρούσα διδακτορική διατριβή, ασχολήθηκε με ένα εύρος θεμάτων και ερευνητικών ερωτημάτων της βιοπληροφορικής, τα οποία είχαν κοινό παρονομαστή την μεθοδολογία υπολογιστικής ανάλυσης βιολογικών δεδομένων και την προσπάθεια εντοπισμού των παραγόντων που εμπλέκονται στην αιτιολογία πολυπαραγοντικών ασθενειών. Ειδικότερα, χρησιμοποιήθηκαν δεδομένα από γονιδιακούς πολυμορφισμούς (SNPs) και η συσχέτιση τους με ασθένειες, δεδομένα γονιδιακής έκφρασης για τη διευρεύνηση του τρόπου που αυτά επηρεάζουν τις ασθένειες, και τέλος, πραγματοποιήθηκε ανάλυση πολύπλοκων βιολογικών δικτύων που αφορούν τη μοριακή βάση ασθενειών, ενσωματώνοντας δεδομένα από πολλαπλές πηγές. Αρχικά, χρησιμοποιήθηκε η μεθοδολογία μετα-ανάλυσης γενετικών δεδομένων για τη διερεύνηση της συσχέτισης γονιδιακών πολυμορφισμών με ασθένειες. Πραγματοποιήθηκε μετα-ανάλυση που είχε ως στόχο τη διερεύνηση των πολυμορφισμών του γονιδίου CD24 με τον κίνδυνο εμφάνισης της σκλήρυνσης κατά πλάκας, μετα-ανάλυση των πολυμορφισμών του γονιδίου FZD3 με τον κίνδυνο εμφάνισης της σχιζοφρένειας και μετα-ανάλυση με στόχο τη διερεύνηση του ρόλου των γενετικών πολυμορφισμών του συστήματος ρενίνης-αγγειοτενσίνης σε νεφρικές ασθένειες. Στη συνέχεια, η μελέτη εστίασε στα δεδομένα γονιδιακής έκφρασης και τον τρόπο που αυτά εμπλέκονται σε ασθένειες. Πραγματοποιήθηκε μια εκτενής ανασκόπηση των στατιστικών μεθόδων που έχουν προταθεί για τον εντοπισμό διαφορικώς εκφρασμένων γονιδίων από ανάλυση και μετα-ανάλυση μικροσυστοιχιών DNA. Υλοποιήθηκαν, δυο μέθοδοι μετα-ανάλυσης βασισμένες στον έλεγχο t-test, ενώ προτάθηκε και μια πρωτότυπη παραλλαγή της μεθόδου η οποία εμφανίζει μια σειρά από πλεονεκτήματα. Οι παραπάνω μέθοδοι εφαρμόστηκαν προκειμένου να βρεθούν γονίδια που εμπλέκονται στο έμφραγμα του μυοκαρδίου. Επόμενος στόχος της διατριβής, ήταν η κατασκευή ενός ενοποιημένου μοντέλου που να χρησιμοποιεί δεδομένα από διαφορετικές πηγές προκειμένου να εξαχθεί ένα συμπέρασμα σχετικά με την εμπλοκή των γονιδίων σε ασθένειες. Χρησιμοποιώντας αξιόπιστα δεδομένα γενετικής συσχέτισης από διαφορετικές πηγές, δημιουργήθηκε ένα δίκτυο αλληλεπιδράσεων ανθρώπινων γονιδίων και ασθενειών, προκειμένου να διερευνηθούν οι συσχετίσεις μεταξύ των γενετικών ασθενειών του ανθρώπου και των γονιδίων που σχετίζονται με τις ασθένειες αυτές. Οι σημαντικές συσχετίσεις ασθενειών–ασθενειών που ανιχνεύθηκαν με κοινή γενετική βάση μπορούν να αποτελέσουν το θεμέλιο λίθο, στον οποίο θα στηριχθούν οι μελλοντικές προσπάθειες. Τέλος, κατασκευάστηκε και μελετήθηκε για πρώτη φορά στη βιβλιογραφία το δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζευγμένων με G-πρωτεΐνες (GPCRs) ώστε να διερευνηθεί η πιθανή συσχέτιση της δράσης των υποδοχέων αυτών με τον κίνδυνο εμφάνισης ασθενειών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Abstract

This PhD thesis dealt with a wide range of topics and research questions of bioinformatics, which had as a common denominator the computational methodology for analysis of biological data and the identification of factors involved in the etiology of multifactorial diseases. In particular, we dealt with data from gene polymorphisms (SNPs) and their association with diseases, with gene expression data and the way they affect disease formation, and finally, with the analysis of complex biological networks regarding the molecular basis of disease, incorporating data from multiple sources. Initially we used the methodologies of meta-analysis of genetic data in order to investigate the association of genetic polymorphisms with diseases. We conducted a meta-analysis in order to investigate the association of CD24 gene polymorphisms with the risk of multiple sclerosis, a meta-analysis of the FZD3 gene polymorphisms and the risk of schizophrenia and a meta-analysis to examine the role of genetic polymorphisms of the renin-angiotensin system in renal diseases. Afterwards, we studied gene expression data and the way they are involved in diseases. We conducted an extensive review of the statistical methods have been proposed for identifying differentially expressed genes for analysis and meta-analysis of microarray DNA. We implemented, two methods of meta-analysis based on t-test, and we proposed a novel variant which exhibits a number of advantages. These methods were employed in order to identify genes involved in myocardial infarction. The next goal of the thesis was to construct a unified model that uses data from different sources in order to draw conclusions on the involvement of genes in diseases. Using reliable genetic association data from different sources, a network of human genes and diseases interactions was created in order to investigate the relationships between the human genetic diseases and the genes associated with these diseases. Major disease-disease correlations detected by common genetic basis could constitute the foundation, on which future efforts could be directed. Finally, we constructed and studied for the first time in the literature the signal transduction system of human G-proteins coupled receptors (GPCRs) in order to investigate the possible association of action of these receptors with the risk of diseases.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Περιεχόμενα

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 1

Εισαγωγή

1.1 Το Γενετικό Υλικό

Το DNA (δεοξυριβονουκλεϊκό οξύ) είναι το γενετικό υλικό σχεδόν όλων των οργανισμών. Ανακαλύφθηκε πρώτη φορά το 1869 ενώ μόλις το 1944 αποδείχθηκε ότι είναι ο φορέας της γενετικής πληροφορίας (Lewin, 1990; Watson, 1987), και το 1953 βρέθηκε η ακριβής μοριακή δομή του. Είναι πολυμερές και τα μονομερή του είναι τα δεοξυριβονουκλεοτίδια, σύνθετα μόρια που αποτελούνται από ένα σάκχαρο, την πεντόζη δεοξυριβόζη, μια φωσφορική ομάδα, και μια αζωτούχο βάση η οποία μπορεί να είναι μια εκ των τεσσάρων Αδενίνη (A), Θυμίνη (T), Γουανίνη (G), ή Κυτοσίνη (C). Επειδή η αζωτούχος βάση χαρακτηρίζει μονοσήμαντα το νουκλεοτίδιο οι όροι «βάση» και «νουκλεοτίδιο» χρησιμοποιούνται χωρίς διάκριση όταν αναφερόμαστε στο DNA (αλλά και στο RNA) (Watson, 1987).

Τα νουκλεοτίδια ενώνονται το ένα με το άλλο με φωσφοδιεστερικό δεσμό. Ο δεσμός αυτός ενώνει το υδροξύλιο του 3΄ άνθρακα της πεντόζης του πρώτου νουκλεοτιδίου με τη φωσφορική ομάδα που βρίσκεται στο 5΄ άνθρακα της πεντόζης του επόμενου νουκλεοτιδίου. Ο δεσμός αυτός είναι 3’→5’ φωσφοδιεστερικός δεσμός και τα νουκλεοτίδια ενώνονται και σχηματίζουν μια αλυσίδα με κατεύθυνση 5’→3’, διότι σε μια πολυνουκλεοτιδική αλυσίδα υπάρχει πάντα ελεύθερη φωσφορική ομάδα συνδεδεμένη στον 5΄ άνθρακα της πεντόζης του πρώτου νουκλεοτιδίου της και το τελευταίο νουκλεοτίδιό της έχει ελεύθερο το υδροξύλιο του 3’ άνθρακα (Alberts et al., 1994).

Το μόριο του DNA σχηματίζει μια δεξιόστροφη διπλή έλικα η οποία αποτελείται από δυο αντιπαράλληλες αλυσίδες (κλώνους). Το βήμα της έλικας είναι 10 ζεύγη βάσεων (σε μήκος 3.4 nm) και το ιδιαίτερο χαρακτηριστικό της, που προσδίδει στο DNA την ιδιότητά του ως γενετικό υλικό, είναι η συμπληρωματικότητα. Συγκεκριμένα, η διπλή έλικα σταθεροποιείται με δεσμούς υδρογόνου μεταξύ των συμπληρωματικών βάσεων, δηλαδή μεταξύ A-T σχηματίζονται δύο δεσμοί υδρογόνου και μεταξύ G-C τρεις δεσμοί (Alberts et al., 1994).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Εικόνα 1.1. Η δομή του γενετικού υλικού DNA.

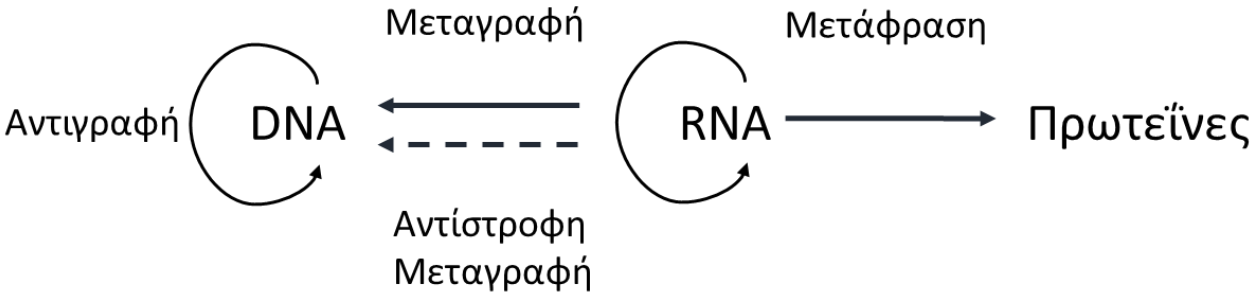
Το RNA (ριβονουκλεϊκό οξύ) είναι το δεύτερο από τα νουκλεϊκά οξέα, περιέχει την αζωτούχο βάση Ουρακίλη (U) αντί της Θυμίνης (T) και έχει για σάκχαρο την πεντόζη ριβόζη αντί της δεοξυριβόζης. Είναι πολύ μικρότερο σε μέγεθος από το DNA και απαντάται σε 4 τύπους: το mRNA, το tRNA, το rRNA και το snRNA. Το mRNA (αγγελιοφόρο RNA) μεταφέρει τη γενετική πληροφορία για τη σύνθεση των πρωτεϊνών, το tRNA (μεταφορικό RNA) μεταφέρει τα απαραίτητα αμινοξέα που θα σχηματίσουν την πρωτεΐνη στα ριβοσώματα, το rRNA (ριβοσωμικό RNA) αποτελεί συστατικό των ριβοσωμάτων και το snRNA (μικρό-πυρηνικό RNA) καταλύει τη διαδικασία της ωρίμανσης του mRNA και συναντάται μόνο στα ευκαρυωτικά κύτταρα (Alberts et al., 1994).

1.2 Το Κεντρικό Δόγμα της Μοριακής Βιολογίας – Γενετικός Κώδικας

Τη λειτουργία του DNA ως γενετικού υλικού των οργανισμών αντικατοπτρίζει τέλεια το λεγόμενο Κεντρικό Δόγμα της Μοριακής Βιολογίας. Σύμφωνα με αυτό, το DNA αντιγράφεται με τη βοήθεια πολλών ενζύμων (DNA πολυμεράση, ελικάση, δεσμάση κλπ) σε μια διαδικασία που χαρακτηρίζεται ως ημισυντηρητικός διπλασιασμός (επειδή κάθε μια από τις δυο αλυσίδες καθορίζει το σχηματισμό μιας συμπληρωματικής της, και τελικά τα δυο θυγατρικά μόρια περιέχουν από μια παλιά και μια

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

νεοσυντιθέμενη αλυσίδα). Η αντιγραφή στα ευκαρυωτικά κύτταρα ξεκινά ταυτόχρονα από πολλά σημεία του γονιδιώματος και προχωρά παράλληλα (Alberts et al., 1994).



Εικόνα 1.2. Το κεντρικό Δόγμα της Βιολογίας. Τα συνεχή βέλη δείχνουν τη ροή της γενετικής πληροφορίας, ενώ το στικτό βέλος απεικονίζει την ειδική περίπτωση παραγωγής DNA από RNA.

Επίσης, κάποια τμήματα του DNA που ονομάζονται γονίδια μεταγράφονται, δηλαδή παράγουν RNA μέσω της διαδικασίας της μεταγραφής. Για τη μεταγραφή απαιτούνται ειδικά ένζυμα (RNA πολυμεράση), ειδικές πρωτεΐνες (μεταγραφικοί παράγοντες) και ειδικές αλληλουχίες πριν από το ενεργό γονίδιο που ονομάζονται υποκινητές. Το στικτό βέλος στην Εικόνα 1.2 υποδηλώνει την αντίστροφη μεταγραφή που συμβαίνει μόνο σε κάποιους ιούς οι οποίοι έχουν RNA για γενετικό υλικό (ρετροϊοί π.χ. HIV).

Το mRNA που παράγεται είναι ο φορέας της γενετικής πληροφορίας, καθώς μεταφέρει το μήνυμα για τη σύνθεση των πρωτεϊνών από το DNA στις πρωτεϊνοσυνθετικές συσκευές των κυττάρων, τα ριβοσώματα. Εκεί με τη συνεργασία των tRNA και rRNA, πραγματοποιείται η μετάφραση, δηλαδή η πληροφορία που υπήρχε στο DNA και μεταφέρθηκε από το RNA μεταφράζεται και χρησιμεύει για την παραγωγή των πρωτεϊνών, οι οποίες είναι μόρια που εξυπηρετούν όλες τις ανάγκες του κυττάρου δομικές και λειτουργικές (ένζυμα κ.α.) (Lewin, 1990).

Κάτι άλλο που πρέπει να τονιστεί είναι το γεγονός ότι στα ευκαρυωτικά κύτταρα το mRNA που παράγεται λέγεται πρόδρομο mRNA και προκειμένου να καταστεί λειτουργικό πρέπει να υποστεί τη διαδικασία της ωρίμανσης η οποία πραγματοποιείται μόνο στους ευκαριωτικούς οργανισμούς. Κατά τη διαδικασία αυτή αποκόπτονται ενδιάμεσα τμήματα τα οποία δεν κωδικοποιούν αμινοξέα, γνωστά ως εσώνια και τα υπόλοιπα τμήματα, τα εξώνια, ενώνονται για να δώσουν το ώριμο mRNA. Η διαδικασία καθορίζεται από ειδικές αλληλουχίες στις άκρες εσωνίων – εξωνίων και καταλύεται από κάποια ριβονουκλεοπρωτεϊνικά σωμάτια (snRNA) (Lewin, 1990).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Η γενετική πληροφορία, η οποία βρίσκεται στο DNA, μεταφέρεται μέσω του RNA και τελικά εκφράζεται με την παραγωγή πρωτεϊνών. Πράγματι η γενετική πληροφορία βρίσκεται στην πρωτοταγή δομή του DNA, δηλαδή στην ακριβή αλληλουχία των βάσεων του, η οποία μεταφέρεται αυτούσια στην αλληλουχία του mRNA. Για να μπορέσει το κύτταρο να αντιστοιχίσει την πληροφορία αυτή με την ακριβή πρωτοταγή δομή των πρωτεϊνών (την ακριβή αλληλουχία των αμινοξέων τους) χρειάζεται έναν κώδικα, το γενετικό κώδικα (Crick et al., 1961).

ΓΕΝΕΤΙΚΟΣ ΚΩΔΙΚΑΣ

		Δεύτερο Γράμμα									
		U		**C**		**A**		**G**			
Πρώτο Γράμμα	**U**	**UUU**	Phe (F)	**UCU**	Ser (S)	**UAU**	Tyr (Y)	**UGU**	Cys (C)	**U**	**Τρίτο Γράμμα**
		UUC	Phe (F)	**UCC**	Ser (S)	**UAC**	Tyr (Y)	**UGC**	Cys (C)	**C**	
		UUA	Leu (L)	**UCA**	Ser (S)	**UAA**	ΛΗΞΗ	**UGA**	ΛΗΞΗ	**A**	
		UUG	Leu (L)	**UCG**	Ser (S)	**UAG**	ΛΗΞΗ	**UGG**	Trp (W)	**G**	
	C	**CUU**	Leu (L)	**CCU**	Pro (P)	**CAU**	His (H)	**CGU**	Arg (R)	**U**	
		CUC	Leu (L)	**CCC**	Pro (P)	**CAC**	His (H)	**CGC**	Arg (R)	**C**	
		CUA	Leu (L)	**CCA**	Pro (P)	**CAA**	Gln (Q)	**CGA**	Arg (R)	**A**	
		CUG	Leu (L)	**CCG**	Pro (P)	**CAG**	Gln (Q)	**CGG**	Arg (R)	**G**	
	A	**AUU**	Ile (I)	**ACU**	Thr (T)	**AAU**	Asn (N)	**AGU**	Ser (S)	**U**	
		AUA	Ile (I)	**ACC**	Thr (T)	**AAC**	Asn (N)	**AGC**	Ser (S)	**C**	
		AUC	Ile (I)	**ACA**	Thr (T)	**AAA**	Lys (K)	**AGA**	Arg (R)	**A**	
		AUG	Met (M)	**ACG**	Thr (T)	**AAG**	Lys (K)	**AGG**	Arg (R)	**G**	
	G	**GUU**	Val (V)	**GCU**	Ala (A)	**GAU**	Asp (D)	**GGU**	Gly (G)	**U**	
		GUA	Val (V)	**GCC**	Ala (A)	**GAC**	Asp (D)	**GGC**	Gly (G)	**C**	
		GUC	Val (V)	**GCA**	Ala (A)	**GAA**	Glu (E)	**GGA**	Gly (G)	**A**	
		GUG	Val (V)	**GCG**	Ala (A)	**GAG**	Glu (E)	**GGG**	Gly (G)	**G**	

Εικόνα 1.3. Ο γενετικός κώδικας.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Ο γενετικός κώδικας είναι ο τρόπος με τον οποίο οι οργανισμοί μεταφράζουν τη «γλώσσα» του DNA με ένα αλφάβητο 4 γραμμάτων (τα 4 νουκλεοτίδια) στη «γλώσσα» των πρωτεϊνών η οποία απαρτίζεται από ένα αλφάβητο 20 γραμμάτων (τα 20 αμινοξέα) (Crick et al., 1961). Όπως είναι φανερό δεν είναι δυνατόν να υπάρχει μια προς μια αντιστοίχιση νουκλεοτιδίων-αμινοξέων και ο μόνος τρόπος να συμβεί αυτό είναι σε κάθε ένα αμινοξύ να αντιστοιχεί μια «λέξη» από 3 ή περισσότερα νουκλεοτίδια. Πράγματι οι δυνατοί συνδυασμοί των 4 νουκλεοτιδίων ανά 3 με επανάληψη είναι 4^3=64 και επαρκούν για την κωδικοποίηση των 20 αμινοξέων, και όντως αυτόν τον τρόπο χρησιμοποιούν οι οργανισμοί (Alberts et al., 1994). Τα γενικά χαρακτηριστικά του γενετικού κώδικα είναι:

- Κώδικας τριπλέτας (3 νουκλεοτίδια –κωδικόνιο-αντιστοιχούν σε ένα αμινοξύ)
- Συνεχής (δεν υπάρχουν νουκλεοτίδια που να μην ανήκουν σε καμία τριπλέτα)
- Μη επικαλυπτόμενος (δεν υπάρχουν νουκλεοτίδια που να ανήκουν σε δυο τριπλέτες)
- Σχεδόν καθολικός (ισχύει σε όλους τους οργανισμούς με ελάχιστες εξαιρέσεις που διαφοροποιούνται 1-2 κωδικόνια)
- Εκφυλισμένος (για 18 από τα 20 αμινοξέα υπάρχουν περισσότερα του ενός κωδικόνια)
- Υπάρχουν κωδικόνια έναρξης (AUG-Met) και λήξης (UGA, UAA, UAG) τα οποία δεν κωδικοποιούν αμινοξέα.

1.3 Πρωτεΐνες

Μία πρωτεΐνη αποτελείται από αμινοξέα (αμινοξικά κατάλοιπα), συνδεδεμένα σε μία γραμμική σειρά η οποία ονομάζεται αμινοξική ακολουθία ή αλληλουχία. Οι πρωτεΐνες σύμφωνα με το κεντρικό δόγμα της Μοριακής Βιολογίας (Crick, 1958; Crick et al., 1961) συντίθενται μονοσήμαντα, με βάση τη γενετική πληροφορία που εμπεριέχεται στο DNA, και η οποία μεταβιβάζεται στην πρωτεϊνοσυνθετική μηχανή των ριβοσωμάτων, μέσω του RNA.

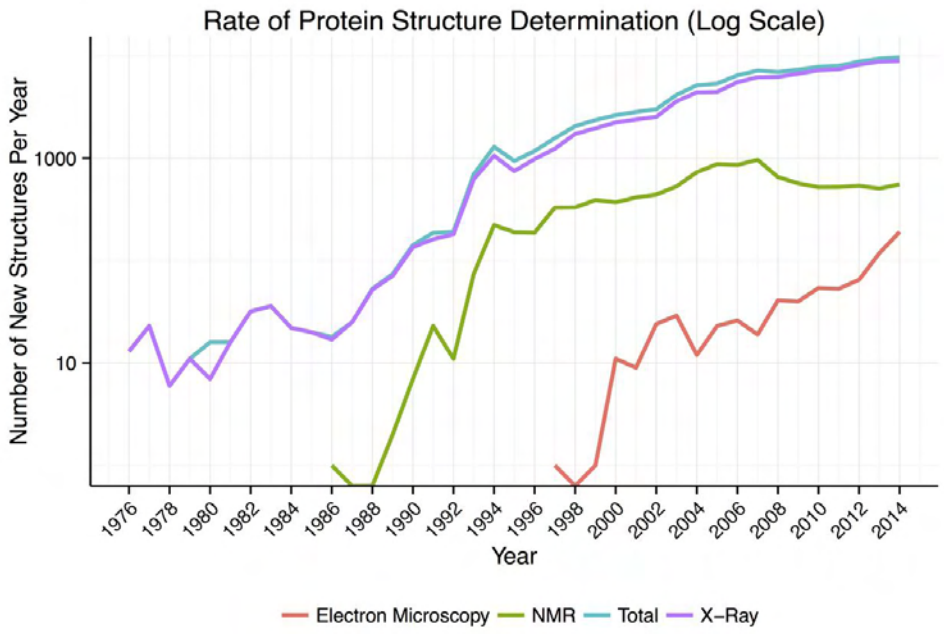
Η βιολογική λειτουργία των πρωτεϊνών εξαρτάται κατά κύριο λόγο από την στερεοδιάταξή τους, δηλαδή από τον τρόπο με τον οποίο η γραμμική αμινοξική τους ακολουθία αναδιπλώνεται στο χώρο (Alberts et al., 1994). Η τρισδιάστατη δομή των πρωτεϊνών έχει βρεθεί με μεθόδους κρυσταλλογραφίας ακτίνων-Χ και φασματοσκοπίας NMR (Berman et al., 2002). Αυτές οι μέθοδοι είναι δαπανηρές, επίπονες και χρονοβόρες και απαιτούν τη χρήση κρυστάλλων (ακτίνες-Χ), οι οποίοι δεν δημιουργούνται εύκολα για αρκετές κατηγορίες πρωτεϊνών. Κατά συνέπεια προέκυψε, η ανάγκη να ληφθούν πληροφορίες για το πρωτεϊνικό δίπλωμα με άλλες μεθόδους. Τις τελευταίες δύο δεκαετίες, οι αμινοξικές ακολουθίες περισσότερων από 150.000 πρωτεϊνών έχουν γίνει γνωστές (Bairoch et al., 2005) και ο

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αριθμός συνεχίζει να αυξάνει με γεωμετρικό ρυθμό. Καθώς τα πειραματικά δεδομένα σαφέστατα δείχνουν ότι όλη η αναγκαία πληροφορία, ώστε μια πρωτεΐνη να διπλωθεί στην φυσική της στερεοδομή, είναι κωδικοποιημένη στην γραμμική αμινοξική της ακολουθία (Anfinsen, 1973), έχουν ξεκινήσει από δεκαετίες προσπάθειες ώστε να προβλεφθεί η δομή μιας πρωτεΐνης από την αμινοξική της ακολουθία (Chou and Fasman, 1978; Rost and Sander, 1993), αλλά με περιορισμένη επιτυχία μόνο.

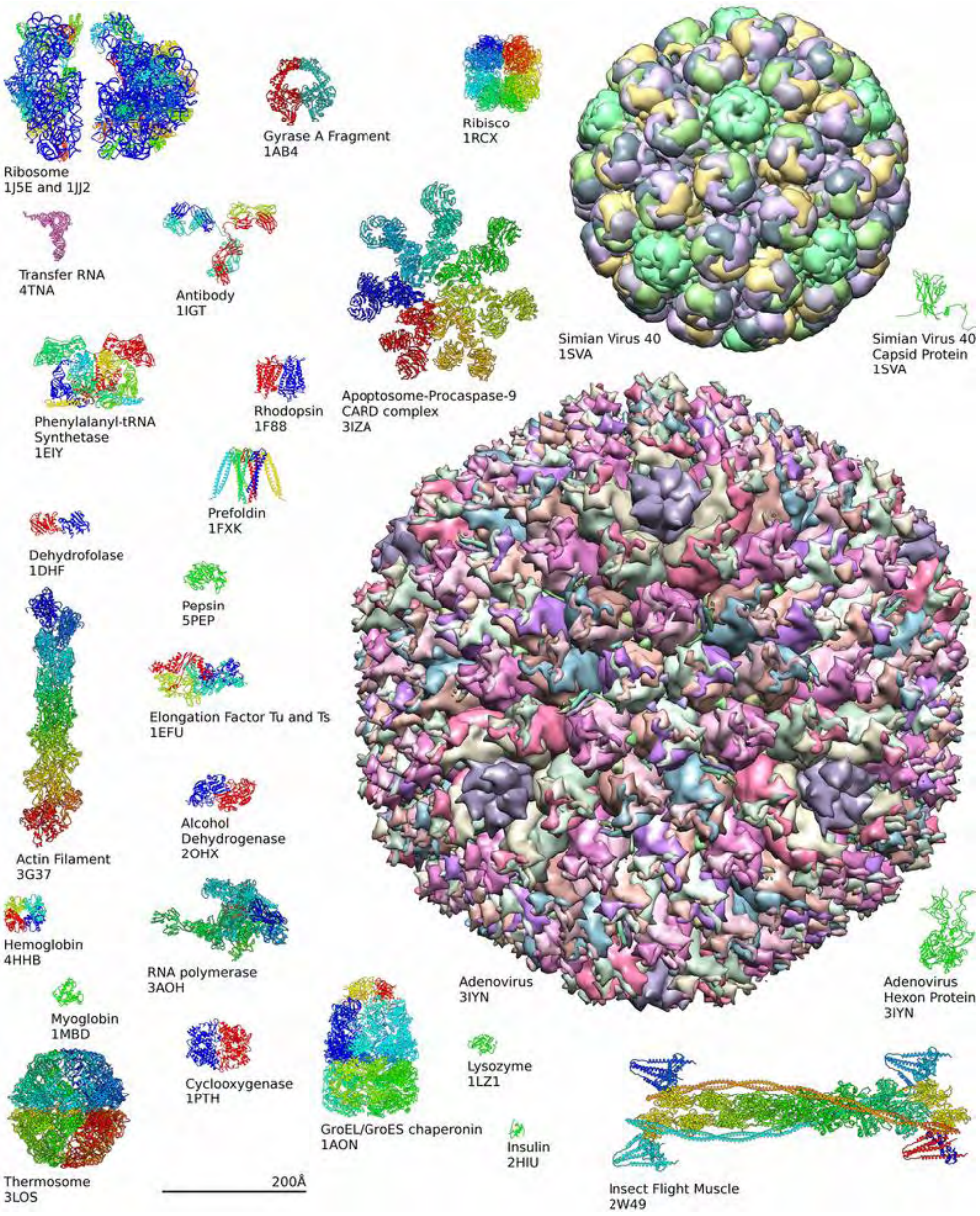
Πρόσφατα, το λεγόμενο πρόγραμμα προσδιορισμού του γονιδιώματος (Human genome project) πολλών οργανισμών μεταξύ των οποίων και του ανθρώπινου κατά συνέπεια και όλων των πρωτεϊνών που αυτό κωδικοποιεί, έχουν ως αποτέλεσμα να συσσωρεύονται τεράστια ποσά πληροφοριών σχετικά με τις ακολουθίες χιλιάδων πρωτεϊνών με ιλιγγιώδεις ρυθμούς. Πρέπει να σημειωθεί εδώ, ότι το πρόβλημα προσδιορισμού της λειτουργίας των πρωτεϊνών εντείνεται καθώς έχει διαπιστωθεί ότι οι ακολουθίες (γονιδίων και πρωτεϊνών) οι οποίες βρίσκονται κατατεθειμένες στις δημόσιες βάσεις δεδομένων αυξάνονται με εκθετικό ρυθμό και διπλασιάζονται κάθε 18-24 μήνες (Μπάγκος, 2015).

Μία πρώτη κατάταξη των πρωτεϊνών θα μπορούσε να διαχωρίσει τις πρωτεΐνες των οργανισμών σε σφαιρικές-υδατοδιαλυτές, σε δομικές, σε μεικτές, δηλαδή πρωτεΐνες με ιδιότητες ενδιάμεσες των δύο προηγούμενων κατηγοριών (δομικές-σφαιρικές) και σε μεμβρανικές πρωτεΐνες (Pasquier et al., 2001). Η τελευταία κατηγορία είναι εξαιρετικά ενδιαφέρουσα για πολλούς λόγους που θα αναφερθούν στο Κεφάλαιο 6.



Εικόνα 1.4. Το ποσοστό των προσδιορισμένων δομών πρωτεϊνών ανά έτος και ανά μέθοδο, (CC BY-SA 4.0, http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 1.5. Παραδείγματα τρισδιάστατων πρωτεϊνικών δομών από τη βάση δεδομένων PDB, (by Axel Griewel - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=32268221).

1.4 Γενετική Ποικιλομορφία

Το γονίδιο είναι ένα τμήμα (ή αλληλουχία) του DNA που κωδικοποιεί ένα μόριο με γνωστή κυτταρική λειτουργία ή διαδικασία. Αλληλόμορφα λέγονται τα γονίδια που βρίσκονται σε όμοιες γενετικές θέσεις των ομόλογων χρωμοσωμάτων και ελέγχουν την ίδια ιδιότητα με ενδεχομένως διαφορετικό τρόπο. Πολλές γενετικές θέσεις χαρακτηρίζονται από τη δημιουργία αλληλομόρφων γονιδίων και έτσι τα

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

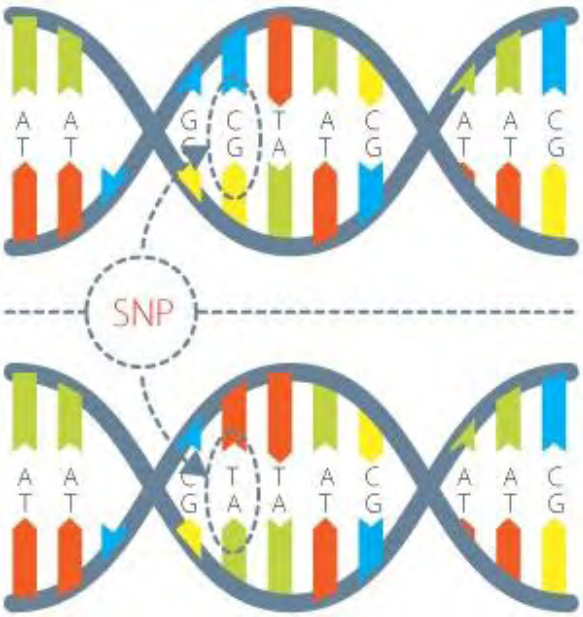
άτομα ενός πληθυσμού εμφανίζουν διαφορετικούς φαινότυπους. Ο γονιδιακός πολυμορφισμός είναι η εμφάνιση πολλαπλών αλληλομόρφων σε μια γενετική θέση στην οποία τουλάχιστον δύο αλληλόμορφα εμφανίζονται με συχνότητα μεγαλύτερη από 1% (Pawson and Linding, 2008).

Οι πολυμορφισμοί δεν είναι κάτι σπάνιο. Ορισμένες φορές προσδίδουν πλεονεκτήματα στον οργανισμό που τους φέρει, σε κάποιες περιπτώσεις οδηγούν σε ασθένεια ή σε προδιάθεση για αυτήν, ενώ σε κάποιες περιπτώσεις δεν έχουν απολύτως καμία επίπτωση. Οι πολυμορφισμοί έχουν μεγάλη σημασία στην ιατρική και χρησιμοποιούνται στη χαρτογράφηση γονιδίων, στην εύρεση γονιδίων που εμπλέκονται με γενετικές ασθένειες, στον έλεγχο πατρότητας και στην εγκληματολογία (Pawson and Linding, 2008).

Οι κυριότερες κατηγορίες γονιδιακών πολυμορφισμών είναι οι πολυμορφισμοί ενός νουκλεοτιδίου ή SNPs που χαρακτηρίζονται από την αντικατάσταση μια βάσης από μια άλλη, οι επαναληπτικές αλληλουχίες (π.χ. το TC_{18} αντικατοπτρίζει 18 επαναλήψεις TC σε μια αλληλουχία) που εμφανίζονται κυρίως στις μη κωδικές αλληλουχίες του γενετικού υλικού και οι προσθήκες ή ελλείψεις διαδοχικών βάσεων που έχουν ως αποτέλεσμα αλλαγή στην αλληλουχία και εμφάνιση διαφορετικού φαινοτύπου. Στην παρούσα διατριβή θα ασχοληθούμε με τους πολυμορφισμούς ενός νουκλεοτιδίου (Nussbaum et al., 2007).

Οι πολυμορφισμοί ενός νουκλεοτιδίου (SNPs) είναι η πιο συνηθισμένη κατηγορία πολυμορφισμών στο ανθρώπινο γονιδίωμα. Ο αριθμός των γνωστών SNPs στο ανθρώπινο γονιδίωμα ξεπερνά τα 10 εκατομμύρια. Οι πολυμορφισμοί ενός νουκλεοτιδίου μπορεί να είναι δύο ειδών: α) πολυμορφισμοί που βρίσκονται σε περιοχές που δεν κωδικοποιούν πρωτεΐνες και έτσι δεν γίνονται αντιληπτοί και β) πολυμορφισμοί σε κωδικοποιές περιοχές όπου είτε αλλάζουν τη λειτουργία της παραγόμενης πρωτεΐνης (μη-συνώνυμη μετάλλαξη) είτε δεν την αλλάζουν (συνώνυμη μετάλλαξη) (Pawson and Linding, 2008). Οι πολυμορφισμοί ενός νουκλεοτιδίου ευθύνονται για τη ύπαρξη της γενετικής ποικιλότητας και χρησιμοποιούνται ως γενετικοί δείκτες για την εύρεση γονιδίων που σχετίζονται με ασθένειες, για τον υπολογισμό του κινδύνου που διατρέχει κάποιος στο να εμφανίσει μια ασθένεια και για τον καλύτερο σχεδιασμό της θεραπευτικής αντιμετώπισης (Nussbaum et al., 2007).

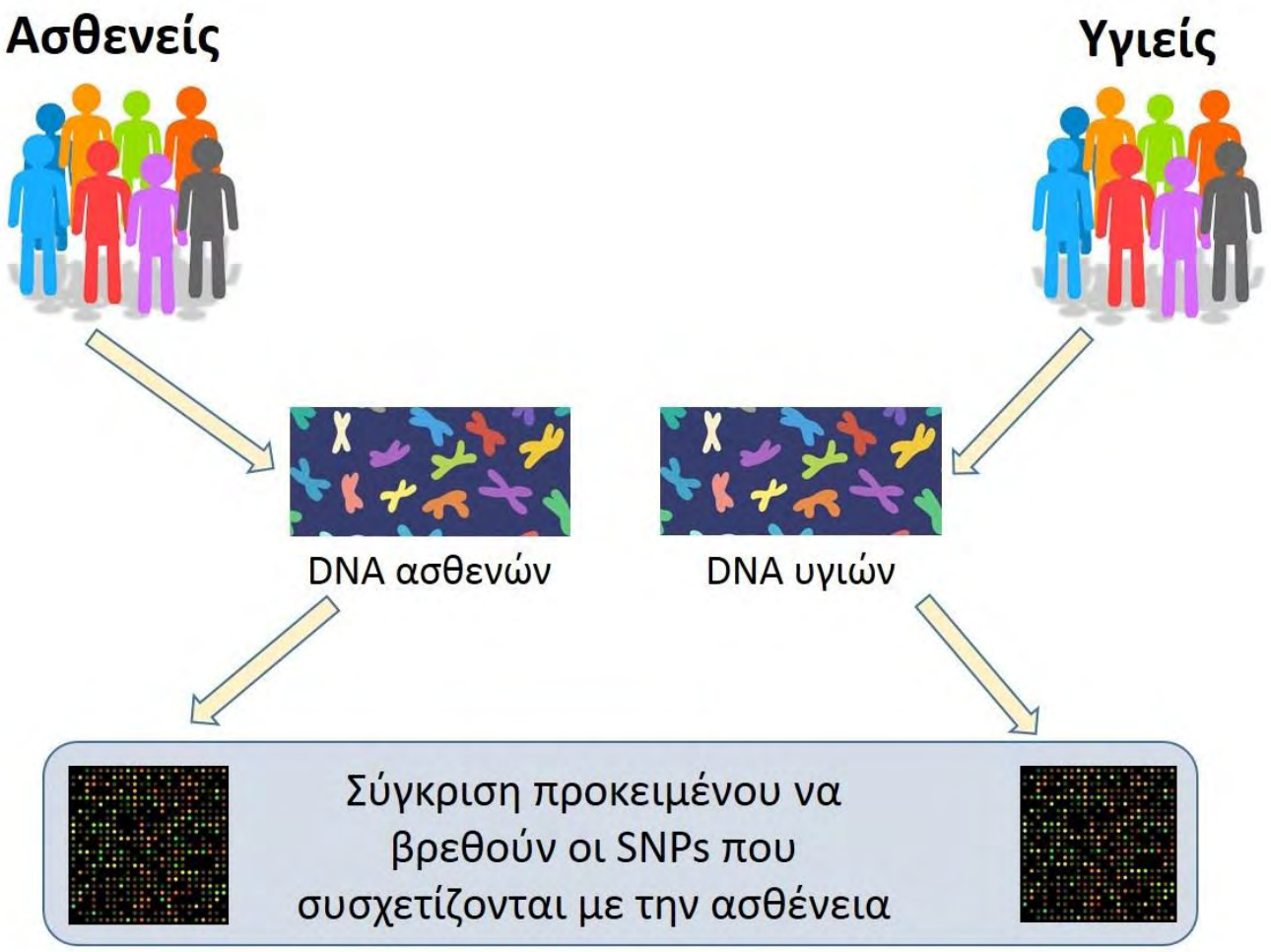
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 1.6. Πολυμορφισμοί ενός νουκλεοτιδίου (SNPs).

Οι μελέτες γενετικής συσχέτισης στοχεύουν στην εύρεση συσχετίσεων μεταξύ ενός ή περισσοτέρων γονιδιακών πολυμορφισμών και ενός χαρακτηριστικού, το οποίο μπορεί να είναι είτε ποσοτικό (αύξηση της τιμής της γλυκόζης στο αίμα) είτε να αποτελεί μια διχοτομική έκβαση (παρουσία-απουσία μιας ασθένειας). Οι μελέτες γενετικής συσχέτισης ελέγχουν μόνο ένα γενετικό πολυμορφισμό και ανιχνεύουν μικρές επιδράσεις και έτσι δημιουργείται η ανάγκη για ταυτόχρονη αξιολόγηση περισσότερων γενετικών δεικτών. Για την εμφάνιση πολύπλοκων ασθενειών ευθύνονται πολλοί γονιδιακοί πολυμορφισμοί που δρουν συνεργατικά, οι περισσότεροι από τους οποίους έχουν μικρές επιδράσεις. Το γεγονός αυτό, σε συνδυασμό με την ανίχνευση μεγάλου αριθμού SNPs σε όλο το γονιδίωμα, έχει οδηγήσει στη μεγάλη σημαντικότητα των μελετών γενετικής συσχέτισης στο χώρο της γενετικής επιδημιολογίας (Livak et al., 1995; Pokorny et al., 1976; Risch and Merikangas, 1996).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 1.7. Διαδικασία μελετών γενετικής συσχέτισης.

1.5 Γονιδιακή Έκφραση

Ο όρος γονιδιακή έκφραση αναφέρεται στην ποσότητα mRNA ή πρωτεϊνών που παράγονται από ένα κύτταρο σε μια δεδομένη στιγμή. Η γονιδιακή έκφραση είναι μια αυστηρά ρυθμιζόμενη διαδικασία και η αρμονική έκφραση ενός μεγάλου αριθμού γονιδίων είναι καθοριστική για τη διατήρηση της σωστής κυτταρικής λειτουργίας. Συνεπώς, η αύξηση ή ελάττωση του επιπέδου έκφρασης συγκεκριμένων γονιδίων ευθύνεται για αρκετές ασθένειες. Η ανάλυση της γονιδιακής έκφρασης βασίζεται στη σύγκριση δειγμάτων. Για παράδειγμα, δειγμάτων από υγιείς ιστούς και από ιστούς ασθενών με στόχο τη μελέτη συγκεκριμένων ασθενειών (Liu et al., 2007a).

Με τη χρήση παραδοσιακών μεθόδων επιτρέπεται η ποσοτική και ποιοτική ανάλυση μόνο ενός μικρού αριθμού γονιδίων κάθε φορά. Για αυτό το λόγο έχουν αναπτυχθεί τεχνολογίες που επιτρέπουν τη

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

γρήγορη και αποτελεσματική ανάλυση ενός πολύ μεγάλου αριθμού γονιδίων. Η τεχνολογία των μικροσυστοιχιών DNA επιτρέπει την ανάλυση της έκφρασης μιας πληθώρας γονιδίων γρήγορα και αποτελεσματικά (Novoradovskaya et al., 2004). Μία από τις πιο διαδεδομένες βιολογικές εφαρμογές αυτής της τεχνολογίας είναι η σύγκριση των επιπέδων έκφρασης ενός συνόλου γονιδίων τα οποία διατηρούνται υπό συγκεκριμένες συνθήκες (κατάσταση Α) με τα επίπεδα έκφρασης ίδιου συνόλου γονιδίων τα οποία προέρχονται από ένα κύτταρο αναφοράς το οποίο διατηρείται υπό φυσιολογικές συνθήκες (κατάσταση Β), π.χ. η σύγκριση υγιών κυττάρων και κυττάρων τα οποία νοσούν, έτσι ώστε να διερευνηθούν οι αιτίες οι οποίες προκαλούν την νόσο. Οι μικροσυστοιχίες βρίσκουν εφαρμογή στην εύρεση λειτουργιών νέων γονιδίων, αφού γονίδια με παρόμοιο τρόπο έκφρασης σε διαφορετικές πειραματικές συνθήκες μπορεί να προσφέρουν χρήσιμες πληροφορίες στον τρόπο δράσης φαρμακευτικών σκευασμάτων και στη διευκόλυνση ταξινόμησης ασθενών (Brazma and Vilo, 2000).

Η σχετικά καινούργια και συνεχώς εξελισσόμενη τεχνολογία Αλληλούχισης Νέας Γενιάς (Next generation sequencing, NGS) αποτελεί μια πιο αξιόπιστη, αποτελεσματική και λιγότερο χρονοβόρα εναλλακτική των μικροσυστοιχιών (Metzker, 2010; Ozsolak and Milos, 2011). Επιτρέπει τον προσδιορισμό της αλληλουχίας (αλληλούχιση) του συνολικού DNA ή cDNA, που προήλθε από το mRNA ενός γονιδίου, με την πραγματοποίηση εκατομμυρίων αντιδράσεων αλληλούχισης μικρών τμημάτων DNA/cDNA ταυτόχρονα. Ο αριθμός των μεταγράφων κάθε γονιδίου χρησιμοποιείται προκειμένου να ελεγχθεί η διαφορική έκφραση του γονιδίου στις διαφορετικές καταστάσεις (ύπαρξη ασθένειας, απόκριση σε θεραπεία κλπ).

1.6 Η Βιοπληροφορική και η Αναγκαιότητα Ανάλυσης Δεδομένων Μεγάλης Κλίμακας

Η Βιοπληροφορική, στο γνωστικό πεδίο της οποίας εντάσσεται αυτή η διατριβή, αποτελεί έναν διεπιστημονικό κλάδο και παρόλο που δεν υπάρχει ένας κοινώς αποδεκτός ορισμός, μια προσπάθεια ορισμού της θα ήταν ο επιστημονικός χώρος όπου η σύμπραξη της Βιολογίας με την Πληροφορική, τη Στατιστική και τα Μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της Βιολογίας. Πρόκειται για γνωστικό χώρο με συγκεκριμένο όσο και ευρύ πεδίο εφαρμογών και αλληλεπίδρασης με τη σύγχρονη δομική, μοριακή, πληθυσμιακή και περιβαλλοντική Βιολογία. Ο κλάδος της Βιοπληροφορικής σήμερα θεωρείται, παγκόσμια, ένας από τους πλέον αναπτυσσόμενους, ενώ έχει ήδη επιδείξει σημαντικά επιτεύγματα και έχει συγκεντρώσει ιδιαίτερα σημαντικές επενδύσεις (Μπάγκος, 2015).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Ο μεγάλος ρυθμός προσδιορισμού των γονιδιωμάτων, οδήγησε στην αλματώδη ανάπτυξη της επιστήμης της γονιδιωματικής και τις διάφορες υποδιαιρέσεις της: τη συγκριτική γονιδιωματική για τη σύγκριση γονιδιωμάτων, τη λειτουργική γονιδιωματική για τις μελέτες γονιδιακής έκφρασης με μικροσυστοιχίες και τη δομική γονιδιωματική για τη μαζική παραγωγή πρωτεϊνών για δομικές μελέτες και κρυσταλλογραφία ακτίνων Χ. Παράλληλα, εμφανίστηκαν και οι τεχνικές αλληλούχισης νέας γενιάς (Next Generation Sequencing), οι οποίες έδωσαν νέα πνοή σε πειράματα γονιδιακής έκφρασης (RNAseq) και έκαναν εύκολο τον εντοπισμό πολυμορφικών θέσεων, ενώ αναμένεται να επηρεάσουν και την προσωποιημένη ιατρική (Μπάγκος, 2015). Με όλα αυτά τα δεδομένα, δημιουργήθηκε μια μεγάλη ανάγκη για ανάπτυξη περισσότερων μεθόδων πρόγνωσης, ανάλυσης και ομαδοποίησης δεδομένων έτσι ώστε να διαχειριστεί αυτός ο τεράστιος όγκος δεδομένων που προέκυψε από τις παραπάνω τεχνικές.

Η γνώση της αλληλουχίας του ανθρώπινου γονιδιώματος, έδωσε επίσης μεγάλη ώθηση στην επιστήμη της Γενετικής Επιδημιολογίας, καθώς με τον εντοπισμό εκατομμυρίων πολυμορφισμών (SNPs) και τη χρήση της τεχνολογίας των GWAS (Genome-Wide Association Studies), δόθηκε η δυνατότητα να κάνουμε μαζικά μελέτες γενετικής συσχέτισης, ελέγχοντας ταυτόχρονα εκατομμύρια πολυμορφισμούς σε μαζική κλίμακα, όπως ακριβώς και με τις μικροσυστοιχίες DNA. Επιπλέον, μεγάλη εξάπλωση έχουν, όπως αναμενόταν, και οι μεθοδολογίες μετα-ανάλυσης και ενοποίησης δεδομένων (data integration). Οι περιοχές αυτές, είναι περιοχές που πλέον η Βιοπληροφορική έρχεται σε επαφή με τη Γενετική και την Επιδημιολογία και τη Βιοστατιστική (Μπάγκος, 2015).

Τα τελευταία χρόνια έχει παρατηρηθεί σημαντική εξέλιξη στη διεξαγωγή πειραμάτων και τη μεγάλη παραγωγή σημαντικών βιολογικών πληροφοριών η οποία οδήγησε στην τεράστια αύξηση του όγκου της πληροφορίας σε όλα τα επίπεδα και ιδιαίτερα στο επίπεδο της ακολουθίας. Στις μέρες μας, έχουν δημιουργηθεί πολλές βάσεις δεδομένων που συγκεντρώνουν την πληροφορία από τις αντίστοιχες πειραματικές διεργασίες ανάλογα με το είδος των δεδομένων που αποθηκεύουν. Οι βάσεις δεδομένων οι οποίες περιέχουν τα βιολογικά δεδομένα όπως αυτά προσδιορίζονται από τους πειραματικούς επιστήμονες, και, συνήθως, περιέχουν επιπλέον ταξινόμηση και σχολιασμό ονομάζονται πρωτογενείς βάσεις δεδομένων. Πρωτογενείς βάσεις δεδομένων είναι οι: Βάσεις δεδομένων ακολουθιών νουκλεοτιδικών/πρωτεινικών ακολουθιών, Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών, Βάσεις δεδομένων γονιδιακής έκφρασης, Βάσεις δεδομένων γενετικής ποικιλομορφίας, Βάσεις δεδομένων βιβλιογραφίας (Μπάγκος, 2015).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Οι βάσεις δεδομένων γενετικής ποικιλομορφίας καταγράφουν τους πολυμορφισμούς και τις συχνότητές τους στους διάφορους πληθυσμούς. Η dbSNP είναι η δημόσια βάση για τους νουκλεοτιδικούς πολυμορφισμούς http://www.ncbi.nlm.nih.gov/snp (Sherry et al., 2001). Εκτός από νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms - SNPs), περιέχει και δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms - DIPs). Η βάση που καταγράφει τις συσχετίσεις των διαφορετικών πολυμορφισμών μεταξύ των ανθρώπινων πληθυσμών είναι η HapMap (http://hapmap.ncbi.nlm.nih.gov/) (HapMap, 2003). Επιπλέον, έχουν δημιουργηθεί βάσεις γενετικών συσχετίσεων οι οποίες περιέχουν πληροφορίες σχετικά με τη συσχέτιση των γονιδίων και των αντίστοιχων πολυμορφισμών τους με διάφορες ασθένειες. Τέτοιες βάσεις είναι: η βάση δεδομένων OMIM (Online Mendelian Inheritance in Man) (Amberger et al., 2015) η οποία περιέχει πληροφορίες για τις γενετικά κληρονομήσιμες ασθένειες, η βάση δεδομένων GAD (Becker et al., 2004) η οποία περιέχει πληροφορίες για πολυπαραγοντικές ασθένειες και η βάση δεδομένων NHRI GWAS (Welter et al., 2014) η οποία περιέχει συσχετίσεις πολυμορφισμών (SNPs) με ασθένειες.

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων οικονομικότερων τσιπ μικροσυστοιχιών αλλά και με την εμφάνιση των τεχνολογιών Next Generation Sequencing, τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό και έτσι υπάρχει ανάγκη αποθήκευσης και ανάλυσης όλων αυτών των δεδομένων. Τη λύση στο παραπάνω πρόβλημα έδωσαν οι βάσεις δεδομένων οι οποίες περιέχουν δεδομένα από χιλιάδες πειράματα γονιδιακής έκφρασης. Οι βάσεις δεδομένων αυτές επιτρέπουν την καταχώρηση αποτελεσμάτων από πειράματα μικροσυστοιχιών και αλληλούχισης νέας γενιάς, ενώ κάποιες από αυτές προσφέρουν και επιπλέον εργαλεία ανάλυσης των δεδομένων αυτών. Οι πιο γνώστες και συχνά χρησιμοποιούμενες βάσεις δεδομένων μικροσυστοιχιών και αλληλούχισης νέας γενιάς (NGS) είναι: 1) η GeneExpression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), η οποία παρέχει δεδομένα γονιδιακής έκφρασης, τόσο από μικροσυστοιχίες όσο και από αλληλούχιση νέας γενιάς (next generation sequenicng) (Barrett and Edgar, 2006) και επιπλέον περιέχει κάποια διαδικτυακά εργαλεία που επιτρέπουν απλές αναλύσεις των δεδομένων της βάσης, 2) η Array Express (http://www.ebi.ac.uk/arrayexpress/), που είναι βάση δεδομένων μικροσυστοιχιών (Brazma et al., 2003) της ίδιας λογικής με την GEO, την οποία περιέχει ως υποσύνολο βάσει της συνεργασίας των ιδρυμάτων και περιέχει επιπλέον εργαλεία για ανάλυση των παρεχόμενων δεδομένων, 3) η Stanford Microarray Database (SMD), που είναι μια βάση δεδομένων που κατασκευάστηκε αρχικά για να καλύπτει τις

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ανάγκες διαμοιρασμού αρχείων των ερευνητών του Stanford, αλλά μετεξελίχθηκε σταδιακά σε ένα δημόσιο αποθετήριο δεδομένων για μικροσυστοιχίες http://smd.stanford.edu/ (Demeter et al., 2007), 4) η RNA Seq Atlas, που είναι βάση δεδομένων πειραμάτων αλληλούχισης νέας γενιάς και είναι διαθέσιμη στην ιστοσελίδα http://medicalgenomics.org/rna_seq_atlas (Krupp et al., 2012). Η εν λόγω βάση περιέχει επιπλέον δεδομένα γονιδιακής έκφρασης από πειράματα μικροσυστοιχιών ενώ παρέχει και επιπλέον εργαλεία ανάλυσης των δεδομένων.

Οι ανθρώπινες ασθένειες αποδίδονται στη συνδυαστική δράση περισσότερων από ένα γονιδίων (Cordell and Clayton, 2005; Goldstein, 2009). Οι μέθοδοι αλληλούχισης μεγάλης-κλίμακας (high-throughput sequencing) επιτρέπουν την ανίχνευση μεγάλου αριθμού γονιδίων που εμπλέκονται σε ασθένειες και εκατομμυρίων μονονουκλεοτιδικών πολυμορφισμών (SNPs) που σχετίζονται με ασθένειες (Hirschhorn, 2009; Manolio, 2010). Δίκτυα που απεικονίζουν συσχετίσεις ασθενειών-γονιδίων επιτρέπουν όχι μόνο τη διερεύνηση της γενετικής πολυπλοκότητας μιας συγκεκριμένης ασθένειας, αλλά και τη σχέση μεταξύ φαινομενικά διαφορετικών ασθενειών (Barabasi et al., 2011; Pawson and Linding, 2008). Επιπλέον, δίκτυα αλληλεπιδράσεων γονιδίων και ασθενειών επιτρέπουν την πρόβλεψη γονιδίων σχετιζόμενων με ασθένειες, την αποσαφήνιση των μηχανισμών που εμπλέκονται στην παθογένεια των ασθενειών και το σχεδιασμό θεραπευτικών στρατηγικών (Barabasi et al., 2011; Pawson and Linding, 2008). Έχει αποδειχθεί ότι τα γονίδια που σχετίζονται με ίδιες ή παρόμοιες ασθένειες έχουν μεγαλύτερη τάση να αλληλεπιδρούν μεταξύ τους (Hartwell et al., 1999; Oti and Brunner, 2007). Γενικότερα, η ανάλυση των πολύπλοκων αλληλεπιδράσεων που εμφανίζονται στα βιολογικά συστήματα, είτε μιλάμε για συσχετίσεις γονιδίων-ασθενειών, είτε για συνέκφραση γονιδίων, είτε για αλληλεπιδράσεις πρωτεϊνών, είτε για ρυθμιστικές αλληλεπιδράσεις μεταγραφικών παραγόντων, αποτελεί ένα αναδυόμενο κλάδο της Βιοπληροφορικής, και η ανάλυση αυτών των πολύπλοκων δικτύων που προκύπτουν από την ενσωμάτωση τέτοιων δεδομένων, έχει δώσει πολύ σημαντικές πληροφορίες στην προσπάθεια κατανόησης των βιολογικών φαινομένων (Vinayagam et al., 2016).

1.7 Σκοπός της Παρούσας Διατριβής

Η παρούσα διδακτορική διατριβή, ασχολήθηκε με ένα εύρος θεμάτων και ερευνητικών ερωτημάτων της Βιοπληροφορικής, τα οποία όλα είχαν κοινό σημείο τη μεθοδολογία υπολογιστικής ανάλυσης βιολογικών δεδομένων και την προσπάθεια εντοπισμού των παραγόντων που εμπλέκονται στην αιτιολογία πολυπαραγοντικών ασθενειών. Ειδικότερα, μελετήθηκαν δεδομένα από γενετικούς πολυμορφισμούς (SNPs) και η συσχέτιση τους με ασθένειες, δεδομένα γονιδιακής έκφρασης και τον

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

τρόπο που αυτά επηρεάζουν τις ασθένειες, και τέλος, πραγματοποιήθηκε ανάλυση πολύπλοκων βιολογικών δικτύων που αφορούν τη μοριακή βάση ασθενειών, ενσωματώνοντας δεδομένα από πολλαπλές πηγές.

Αρχικός στόχος της παρούσας διατριβής ήταν η διερεύνηση της συσχέτισης γονιδιακών πολυμορφισμών με ασθένειες. Για πολλές γενετικές συσχετίσεις υπάρχει μεγάλος αριθμός μελετών οι οποίες όμως σε πολλές περιπτώσεις αποκλίνουν και έτσι δεν μπορεί να εξαχθεί ένα τελικό συμπέρασμα που να αφορά τη συσχέτιση του πολυμορφισμού. Για το λόγο αυτό χρησιμοποιείται η μεθοδολογία της μετα-ανάλυσης η οποία συνδυάζει τα αποτελέσματα των επιμέρους μελετών που ερευνούν την ίδια συσχέτιση προκειμένου να εξαχθεί ένα τελικό αποτέλεσμα. Στην παρούσα διατριβή πραγματοποιήθηκαν τρεις μετα-αναλύσεις γενετικών δεδομένων. Η πρώτη μετα-ανάλυση είχε ως στόχο τη διερεύνηση των πολυμορφισμών του γονιδίου CD24 με τον κίνδυνο εμφάνισης της σκλήρυνσης κατά πλάκας. Πραγματοποιήθηκε μετα-ανάλυση σε τέσσερις πολυμορφισμούς του γονιδίου CD24. Στη συνέχεια, πραγματοποιήθηκε μετα-ανάλυση των πολυμορφισμών του γονιδίου FZD3 με τον κίνδυνο εμφάνισης της σχιζοφρένειας. Συνολικά διερευνήθηκαν έξι πολυμορφισμοί. Τέλος, πραγματοποιήθηκε μια τρίτη μετα-ανάλυση με στόχο τη διερεύνηση του ρόλου των γενετικών πολυμορφισμών του συστήματος ρενίνης-αγγειοτενσίνης σε νεφρικές ασθένειες.

Επόμενος στόχος της διατριβής ήταν η χρησιμοποίηση δεδομένων γονιδιακής έκφρασης προκειμένου να βρεθούν συσχετίσεις με ασθένειες. Λόγω του μεγάλο όγκου γονιδιακών δεδομένων και της ετερογένειας που υπάρχει μεταξύ των μεθόδων παραγωγής τους υπάρχει η ανάγκη για τη δημιουργία νέων μεθόδων ανάλυσης και μετα-ανάλυσης των δεδομένων αυτών. Έτσι, πραγματοποιήθηκε μια μελέτη ανασκόπησης των μεθόδων που χρησιμοποιούνται για ανάλυση δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA και στη συνέχεια υλοποιήθηκαν στο στατιστικό πακέτο STATA δυο στατιστικές μέθοδοι ανάλυσης δεδομένων μικροσυστοιχιών βασισμένες στον έλεγχο t-test. Η υλοποίηση των μεθόδων είναι διαθέσιμη στο ευρύ κοινό στην ιστοσελίδα: http://www.compgen.org/tools/microarrays#TOC-Methods-of-analysis.

Στη συνέχεια, πραγματοποιήθηκε ανασκόπηση σχετικά με τη μεθοδολογία που ακολουθείται για τη μετα-ανάλυση δεδομένων γονιδιακής έκφρασης και υλοποιήθηκαν δυο μέθοδοι μετα-ανάλυσης βασισμένες στον έλεγχο t-test, ενώ προτάθηκε και μια καινούργια παραλλαγή της μεθόδου η οποία εμφανίζει μια σειρά από πλεονεκτήματα. Η υλοποίηση των μεθόδων είναι διαθέσιμη στο ευρύ κοινό στην ιστοσελίδα: http://www.compgen.org/tools/microarrays#TOC-Methods-of-meta-analysis. Τέλος, οι παραπάνω μέθοδοι εφαρμόστηκαν προκειμένου να βρεθούν γονίδια που εμπλέκονται στο έμφραγμα του

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

μυοκαρδίου. Πραγματοποιήθηκε ανάλυση και μετα-ανάλυση σε 31180 γονίδια από τέσσερις μελέτες μικροσυστοιχιών DNA προκειμένου να βρεθούν ποια είναι αυτά που υπερ ή υπο εκφράζονται στο έμφραγμα του μυοκαρδίου.

Επιπλέον, στόχος της διατριβής ήταν η κατασκευή ενός ενοποιημένου μοντέλου που να χρησιμοποιεί δεδομένα από διαφορετικές πηγές προκειμένου να εξαχθεί ένα συμπέρασμα σχετικά με την εμπλοκή των γονιδίων σε ασθένειες. Χρησιμοποιώντας γενετικά δεδομένα από διαφορετικές πηγές, δημιουργήθηκε ένα δίκτυο αλληλεπιδράσεων ανθρώπινων γονιδίων και ασθενειών, προκειμένου να διερευνηθούν οι συσχετίσεις μεταξύ των γενετικών ασθενειών του ανθρώπου και των γονιδίων που σχετίζονται με τις ασθένειες αυτές. Στη συνέχεια, με βάση το δίκτυο αυτό δημιουργήθηκε ένα δίκτυο γονιδίων και ένα δίκτυο ασθενειών με στόχο την εύρεση ασθενειών που προκαλούνται από κοινά γονίδια και γονιδίων που εμπλέκονται στην ίδια ασθένεια. Η ανάλυση των δικτύων αυτών, έδωσε αρκετά σημαντικά συμπεράσματα σχετικά με τη φύση των πολυπαραγοντικών ασθενειών και την εμπλοκή των γονιδίων σε αυτές.

Τέλος, μελετήθηκε για πρώτη φορά το δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζευγμένων με G-πρωτεΐνες (GPCRs) ώστε να διερευνηθεί η πιθανή συσχέτιση της δράσης των υποδοχέων αυτών με τον κίνδυνο εμφάνισης ασθενειών. Τα γονίδια που κωδικοποιούν τους GPCRs έχουν τεράστια φαρμακολογική σημασία, με περίπου το 40% όλων των φαρμάκων που κυκλοφορούν στην αγορά να στοχεύουν αυτές τις πρωτεΐνες (Hopkins and Groom, 2002; Overington et al., 2006). Τα μονοπάτια μεταγωγής σήματος των GPCR εμπλέκονται σε μια πληθώρα ασθενειών (Dorsam and Gutkind, 2007; Thompson et al., 2014; Vischer et al., 2014). Ως εκ τούτου, η αποσαφήνιση των μοριακών μηχανισμών του μονοπατιού της μεταγωγής σήματος των GPCR καθώς και ο τρόπος με τον οποίο το μονοπάτι αυτό συνδέεται με ασθένειες είναι υψίστης βιολογικής και φαρμακευτικής σημασίας. Η ανάλυση του δικτύου, έδωσε σημαντικές πληροφορίες για τον τρόπο δράσης των υποδοχέων αυτών αλλά και για τον τρόπο που αυτοί επηρεάζουν την εμφάνιση γνωστών ασθενειών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 2

Μετα-ανάλυση Γενετικών Δεδομένων

Η συνεχής αύξηση των μελετών γενετικής συσχέτισης, των ευρυγονιδιωματικών μελετών και των μελετών γονιδιακής έκφρασης καθιστά επιτακτική την ανάγκη του συνδυασμού και της περαιτέρω ανάλυσης των παραπάνω δεδομένων. Η μετα-ανάλυση είναι ένα στατιστικό εργαλείο το οποίο επεξεργάζεται τα δεδομένα και τα αποτελέσματα μελετών, που ερευνούν το ίδιο ερώτημα. Οι αρχές της μεθόδου της μετα-ανάλυσης ξεκινούν από τη εποχή του Fisher, το 1920. Η ουσιαστική μέθοδος της μετα-ανάλυσης, όμως, αναπτύχθηκε από τον Eugene Glass, το 1976, και παρέχει ένα τελικό συμπέρασμα το οποίο προέρχεται από μια σύνθεση ανεξάρτητων δεδομένων.

Για τη διεξαγωγή της μετα-ανάλυσης θα πρέπει πρώτα να καθοριστεί το αντικείμενο μελέτης και στη συνέχεια θα πρέπει να γίνει εκτενής αναζήτηση στη βιβλιογραφία με σκοπό την εύρεση όλων των διαθέσιμων μελετών (Normand, 1999). Μόλις συγκεντρωθεί ο αριθμός των μελετών, εξάγονται τα δεδομένα τα οποία θα χρησιμοποιηθούν για τη μετα-ανάλυση. Στη συνέχεια, καθορίζεται το μέγεθος επίδρασης, το οποίο μπορεί να είναι η διαφορά των μέσων τιμών, ο σχετικός κίνδυνος (relative risk) και ο σχετικός λόγος συμπληρωματικών πιθανοτήτων (Odds Ratio-OR). Η δύναμη της μετα-ανάλυσης εξαρτάται από τον αριθμό των μελετών και από τη μέθοδο την οποία χρησιμοποιούμε για να συνδυάσουμε τις μεμονωμένες εκτιμήσεις του μεγέθους επίδρασης οι οποίες προέρχονται από τις αρχικές μελέτες. Δύο μοντέλα χρησιμοποιούνται για το συνδυασμό των παραπάνω εκτιμήσεων: το μοντέλο των σταθερών επιδράσεων (fixed effect model) και το μοντέλο των τυχαίων επιδράσεων (random effect model).

2.1 Μέθοδοι Μετα-ανάλυσης

Για να διεξαχθεί σωστά μια μετα-ανάλυση θα πρέπει πρώτα να συγκεντρωθούν όλες οι μελέτες που αφορούν το θέμα το οποίο μελετάται και στη συνέχεια να προσδιοριστεί το μέγεθος επίδρασης θ, ο παράγοντας δηλαδή με τον οποίο θα γίνει η μετα-ανάλυση και η διακύμανση s_i^2 της κάθε μελέτης.

Στα συνεχή δεδομένα, ως μέγεθος επίδρασης θ χρησιμοποιείται η διαφορά των μέσων τιμών μεταξύ των δειγμάτων ελέγχου και αναφοράς. Η διαφορά των μέσων τιμών η οποία χρησιμοποιείται ως μέγεθος επίδρασης μπορεί να είναι τυποποιημένη (standardized mean difference) ή μη τυποποιημένη (un standardized mean difference) (Thakkinstian et al., 2005). Στον πίνακα 2.1 παρουσιάζεται η μορφή

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

των δεδομένων μιας μετα-ανάλυσης μελετών γενετικής συσχέτισης με μέγεθος επίδρασης ένα συνεχές χαρακτηριστικό μεταξύ των δυο ομάδων (υγιείς και ασθενείς).

Πίνακας 2.1. Παράδειγμα μιας μορφής δεδομένων που χρησιμοποιούνται για μετα-ανάλυση δεδομένων γενετικής συσχέτισης με μέγεθος επίδρασης ένα συνεχές χαρακτηριστικό μεταξύ δυο ομάδων (υγιείς και ασθενείς).

	Υγιείς		**Ασθενείς**	
Μελέτη	**Μέση Τιμή (Τυπική Απόκλιση-sd)**	**Αριθμός Ατόμων**	**Μέση Τιμή (Τυπική Απόκλιση-sd)**	**Αριθμός Ατόμων**
1	x_{11} (sd_{11})	n_{11}	x_{21} (sd_{21})	n_{21}
2	x_{12} (sd_{12})	n_{12}	x_{22} (sd_{22})	n_{22}
…	…	…	…	…
i	x_{1i} (sd_{1i})	n_{1i}	x_{2i} (sd_{2i})	n_{2i}

Η τυποποιημένη διάφορα μέσων τιμών που χρησιμοποιείται ως μέγεθος επίδρασης με την αντίστοιχη τυπική απόκλιση δίνεται από τις παρακάτω εξισώσεις:

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{sd_i} \quad \text{με} \quad sd_i = \sqrt{\frac{(n_{1i}-1)sd_{1i}^2 + (n_{2i}-1)sd_{2i}^2}{n_{1i}+n_{2i}-2}}$$

Η μη τυποποιημένη διάφορα μέσων τιμών δίνεται από την εξίσωση:

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

Στα διακριτά δεδομένα ως μέγεθος επίδρασης χρησιμοποιείται ο σχετικός λόγος συμπληρωματικών πιθανοτήτων (Odds Ratio - OR). Για παράδειγμα, έστω ότι πρέπει να διερευνηθεί η επίδραση ενός φαρμάκου σε μια ασθένεια. Το Odds Ratio σε αυτή την περίπτωση υπολογίζεται:

$$\text{OR} = \frac{\text{ποσοστό ασθενών που τους χορηγήθηκε το φάρμακο/ποσοστό ασθενών που δεν έλαβαν το φάρμακο}}{\text{ποσοστό υγιών που τους χορηγήθηκε το φάρμακο/ποσοστό υγιών που δεν έλαβαν το φάρμακο}}$$

Στον Πίνακα 2.2 απεικονίζεται μια μορφή δεδομένων που χρησιμοποιούνται για μετα-ανάλυση δεδομένων γενετικής συσχέτισης με τη χρήση του σχετικού λόγου συμπληρωματικών πιθανοτήτων.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.2. Μορφή δεδομένων που χρησιμοποιούνται για μετα-ανάλυση δεδομένων γενετικής συσχέτισης με τη χρήση του σχετικού λόγου συμπληρωματικών πιθανοτήτων.

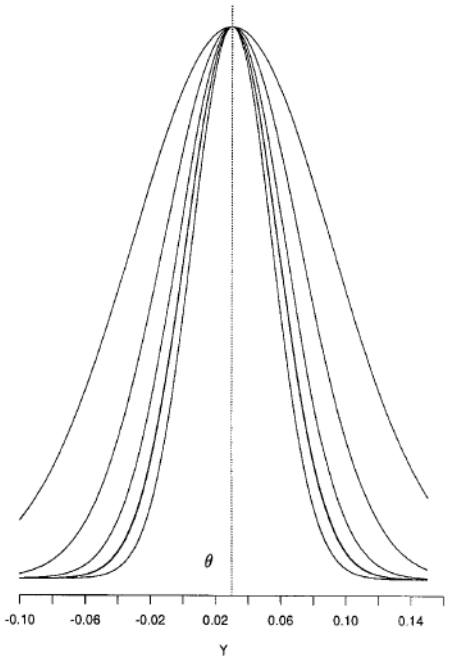
	Ασθένεια (+)	**Ασθένεια (-)**
Φάρμακο (+)	α	β
Φάρμακο (-)	γ	δ

Σύμφωνα με τον παραπάνω πίνακα ισχύει ότι:

$$OR = \frac{\frac{\alpha}{\gamma}}{\frac{\beta}{\delta}}$$

Για το συνδυασμό των μεγεθών επίδρασης των επιμέρους μελετών χρησιμοποιούνται δύο είδη μοντέλων στατιστικής ανάλυσης: το μοντέλο σταθερών επιδράσεων και το μοντέλο τυχαίων επιδράσεων (Normand, 1999). Το μοντέλο σταθερών επιδράσεων υποθέτει ότι όλα τα δείγματα των μελετών προέρχονται από έναν ενιαίο πληθυσμό που έχει ένα κοινό μέγεθος επίδρασης θ (Εικόνα 2.1). Ο τύπος του μοντέλου εκφράζεται ως εξής:

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



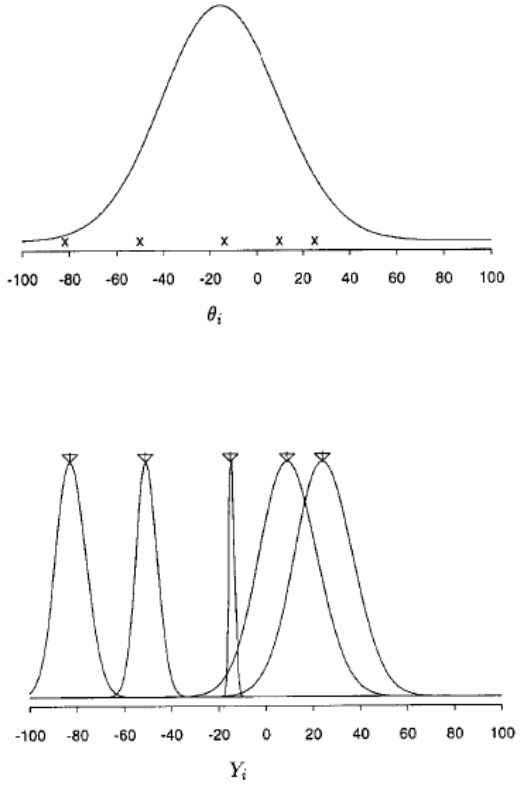
Εικόνα 2.1. Μοντέλο σταθερών επιδράσεων. Η κατανομή πέντε υποθετικών δειγμάτων χρησιμοποιώντας το μοντέλο. Κάθε δείγμα Y_i έχει ένα κοινό μέγεθος επίδρασης θ. Η διαφορά ανάμεσα στις πέντε μελέτες είναι η διαφορετική διακύμανση κάθε μελέτης s_i^2, δηλαδή πόσο καλά υπολογίζει κάθε μελέτη το θ (Normand, 1999).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Το μοντέλο τυχαίων επιδράσεων υποθέτει ότι τα δείγματα της μελέτης τα οποία περιλαμβάνονται σε μια μετα-ανάλυση μπορούν να προέλθουν από μια κατανομή πληθυσμών, δηλαδή επιτρέπει την ύπαρξη ετερογένειας. Κάθε μελέτη έχει και ένα διαφορετικό μέγεθος επίδρασης θ_i και διακύμανση s_i^2 με τύπο $\mathrm{Y}_i \mid \theta_i, s_i^2 \sim \mathrm{N}\left(\theta_i, s_i^2\right)$ (Εικόνα 2.2). Κάθε δείγμα του υπερπληθυσμού έχει μέγεθος επίδρασης το οποίο κατανέμεται με μέση τιμή θ και διακύμανση τ^2, με τύπο $\theta_i \mid \theta, \tau^2 \sim \mathrm{N}\left(\theta, \tau^2\right)$, όπου θ και τ^2 οι υπερπαράμετροι οι οποίες αντιπροσωπεύουν το κοινό μέγεθος επίδρασης και τη διακύμανση, αντίστοιχα. Η ανάλυση του υπερπληθυσμού δίνεται από τον τύπο:

$$\theta_i \mid \mathrm{y}, \theta, \tau^2 \sim \mathrm{N}\left(\mathrm{B}_i\theta + (1-\mathrm{B}_i)Y_i, \mathrm{s}_i^2(1-\mathrm{B}_i)\right)$$

με $y = (Y_1, Y_2, ..., Y_k)$ και το B_i ορίζεται ως $\frac{1}{s_i^2 + \tau^2}$. Όταν το $\tau^2 = 0$ τότε το μοντέλο τυχαίων επιδράσεων είναι ισοδύναμο με το μοντέλο σταθερών επιδράσεων.



Εικόνα 2.2. Μοντέλο τυχαίων επιδράσεων. Κατανομή πέντε υποθετικών δειγμάτων κάνοντας χρήση μοντέλου τυχαίων επιδράσεων. Κάθε μέγεθος επίδρασης θ_i προέρχεται από τον υπερπληθυσμό με μέγεθος επίδρασης θ και διακύμανση τ^2. Στο παράδειγμα κάθε ένα από τα μεγέθη επίδρασης δημιουργήθηκε από τα πέντε αποτελέσματα των μελετών (Normand, 1999).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Η διεξαγωγή μιας σωστής μετα-ανάλυσης προϋποθέτει όσο το δυνατόν εκτενέστερη αναζήτηση στη βιβλιογραφία, να βρεθούν δηλαδή όλες οι διαθέσιμες μελέτες. Κατά καιρούς έχουν εμφανισθεί διάφορα προβλήματα βιβλιογραφίας (γκρίζα βιβλιογραφία, ξενόγλωσση βιβλιογραφία, συστηματικό σφάλμα δημοσίευσης, φαινόμενο του Πρωτέα) τα οποία αν δεν ληφθούν υπόψη, μπορούν να οδηγήσουν σε υπερεκτίμηση των αποτελεσμάτων της μετα-ανάλυσης (Normand, 1999). Επιπλέον, ένα κρίσιμο στάδιο κατά τη διεξαγωγή μιας μετα-ανάλυσης είναι η εξέταση του βαθμού ετερογένειας μεταξύ των μεμονωμένων μελετών. Η ανομοιογένεια μεταξύ των δειγμάτων που εξετάζονται μπορεί να επηρεάσει τα αποτελέσματα της μετα-ανάλυσης και να οδηγήσει και αυτή σε υπερεκτίμηση ή υποεκτίμηση εν προκειμένω του τελικού αποτελέσματος (Thakkinstian et al., 2005).

2.2 Έλεγχος Ετερογένειας

Όταν χρησιμοποιείται το μοντέλο σταθερών επιδράσεων υπάρχει κίνδυνος να υπάρχει ετερογένεια στη μελέτη, επειδή τα δείγματα μπορεί να προέρχονται από την ύπαρξη διαφορετικών πληθυσμών. Η παρουσία ετερογένειας μεταξύ των μελετών μπορεί να προσδιοριστεί από τον έλεγχο Q του Cochran. Το Q δίνεται από τον ακόλουθο τύπο:

$$Q = \sum_i^k Wi(d_i - D)^2 \sim \chi_{k-1}^2$$

$$\text{με} \qquad D = \frac{\sum_{i=1}^k w_i d_i}{\sum_{i=1}^k w_i} \qquad \text{και} \qquad w_i = \frac{1}{\text{var}(d_i)}$$

Το Q ακολουθεί την κατανομή χ^2 με k-1 βαθμούς ελευθερίας όπου k ο αριθμός των μελετών και μπορεί να εφαρμοστεί για τον έλεγχο της ετερογένειας τόσο σε συνεχή όσο και σε διακριτά δεδομένα. Ο έλεγχος Q του Cochran αποδίδει καλύτερα όταν η μετα-ανάλυση περιλαμβάνει πολλές μελέτες.

Μια άλλη στατιστική μέθοδος η οποία χρησιμοποιείται για τον υπολογισμό της ετερογένειας είναι ο έλεγχος I^2 ο οποίος βασίζεται στον έλεγχο Q. Ο I^2 υπολογίζεται από τον ακόλουθο τύπο:

$$I^2 = \max\left(0, \frac{Q-(k-1)}{Q}\right)$$

και παίρνει τιμές από 0 έως 100%. Τιμές κάτω από 25% φανερώνουν μια μικρή ή αμελητέα ετερογένεια, ενώ τιμές οι οποίες υπερβαίνουν το 50% υποδηλώνουν ότι η ετερογένεια αποτελεί σοβαρό πρόβλημα.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Ένας τρίτος εκτιμητής για την ύπαρξη ετερογένειας είναι το τ^2 το οποίο υπολογίζει τη μεταβλητότητα μεταξύ των δύο μελετών. Ο εκτιμητής τ^2 ορίστηκε αρχικά από τους Dersimonian και Laird με τον ακόλουθο τύπο (Normand, 1999), αλλά υπάρχουν και άλλες μέθοδοι εκτίμησης:

$$\tau_{DL}^2=\max\left\{0,\frac{Qw-(k-1)}{\sum Wi-\frac{\sum w_i^2}{\sum w_i}}\right\}$$

Υψηλές τιμές των δεικτών αυτών φανερώνουν την ύπαρξη ετερογένειας ενώ όταν είναι ίσοι ή πλησιάζουν το μηδέν συμπεραίνουμε ότι δεν υπάρχει ετερογένεια μεταξύ των μελετών.

Για να εξεταστεί η ύπαρξη ετερογένειας θεωρούνται οι παρακάτω υποθέσεις:

- H0: Τα δείγματα είναι ομοιογενή
- H1: Τα δείγματα είναι ανομοιογενή

Χρησιμοποιώντας το μοντέλο τυχαίων επιδράσεων, οι τύποι για τον υπολογισμό της ετερογένειας είναι:

$$\theta\,(\tau)_{MLE}=\frac{\sum_i Wi(\tau)Yi}{\sum_k Wi(\tau)} \quad \text{με} \quad W_i\,(\tau)=\frac{1}{s_i^{\,2}+\tau^2} \quad \text{και} \quad Y_i=\bar{x}_{1i}-\bar{x}_{2i}$$

Αν η H0 απορριφθεί, τότε υπάρχει ετερογένεια μεταξύ των μελετών.

2.3 Προβλήματα βιβλιογραφίας

Η διεξαγωγή μιας σωστής μετα-ανάλυσης όπως αναφέραμε και παραπάνω προϋποθέτει όσο το δυνατόν καλύτερη και εκτενέστερη αναζήτηση στη βιβλιογραφία. Κατά την αναζήτηση της βιβλιογραφίας προκύπτουν κάποια προβλήματα τα οποία πρέπει να ληφθούν υπόψη έτσι ώστε να μην οδηγηθούμε σε εσφαλμένη εκτίμηση τον αποτελεσμάτων μας. Το πρώτο πρόβλημα το οποίο πρέπει να ληφθεί υπόψη είναι το φαινόμενο της «γκρίζας» βιβλιογραφίας (grey literature) στην οποία εντάσσονται οι μελέτες οι οποίες δεν έχουν δημοσιευθεί σε κάποιο περιοδικό επειδή πιθανώς παρέχουν αρνητικά αποτελέσματα. Το σφάλμα αυτό προκύπτει από το γεγονός ότι μελέτες οι οποίες καταφέρνουν να βρουν μια συσχέτιση δημοσιεύονται πιο γρήγορα και πιο εύκολα από ότι μελέτες οι οποίες δεν βρίσκουν συσχέτιση. Ένα ακόμη πρόβλημα είναι το φαινόμενο της «ξενόγλωσσης» βιβλιογραφίας σύμφωνα με το οποίο πολλές μελέτες που διεξάγονται σε μη αγγλόφωνες χώρες (Κίνα, Ρωσία) δημοσιεύουν τα αποτελέσματα τους σε τοπικά περιοδικά τα οποία δεν ανήκουν σε διεθνείς βάσεις δεδομένων με αποτέλεσμα κατά την

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αναζήτηση να χάνονται δεδομένα από αυτές τις μελέτες. Τα δύο φαινόμενα τα οποία περιγράφηκαν παραπάνω αποτελούν τις περιπτώσεις του συστηματικού σφάλματος δημοσίευσης (publication bias) (Egger et al., 1997).

2.4 Μετα-ανάλυση των πολυμορφισμών του γονιδίου CD24 με τον κίνδυνο εμφάνισης σκλήρυνσης κατά πλάκας

2.4.1 Εισαγωγή

Η Σκλήρυνση κατά Πλάκας (ΣκΠ) είναι μια χρόνια νόσος του κεντρικού νευρικού συστήματος (ΚΝΣ) με επιπολασμό περίπου 0,1% στους ενήλικές νεαρούς καυκάσιους. Η συνηθέστερη ηλικία εμφάνισης της νόσου είναι μεταξύ 20 και 40, με δύο φορές μεγαλύτερη συχνότητα εμφάνισης στις γυναίκες (Ebers et al., 1995). Γίνεται όλο και περισσότερο σαφές ότι οι περιβαλλοντικοί παράγοντες δρουν σε συνεργασία με τη γενετική προδιάθεση ώστε να εμφανιστεί η ΣκΠ (Ebers, 2008). Ωστόσο, υπάρχουν μερικά γονίδια που έχουν πρόσφατα μελετηθεί για τη συσχέτισή τους με την αιτιοπαθογένεια της νόσου, όπως για παράδειγμα τα γονίδια IL2, IL7 και CD24 (International Multiple Sclerosis Genetics et al., 2007). Το γονίδιο CD24 βρίσκεται στην περιοχή 6q21 (Liu and Zheng, 2007). Η πρωτεΐνη CD24 είναι μια GPI-γλυκοπρωτεΐνη της επιφανείας του κυττάρου και εκφράζεται σε αφθονία σε μια ποικιλία τόσο αιμοποιητικών κυττάρων (κύτταρα Τ, κύτταρα Β, μακροφάγα, ουδετερόφιλα) όσο και σε κύτταρα του ΚΝΣ (αστροκύτταρα και μικρογλοία) που εμπλέκονται στην παθογένεια της ΣκΠ (Hernandez-Campo et al., 2007; Zhou et al., 2003). Η πρωτεΐνη CD24 είναι υπεύθυνη για τον τοπικό πολλαπλασιασμό των Τ κυττάρων, μετά τη μετανάστευσή τους προς το ΚΝΣ, καθώς και για την πρόοδο της πειραματικής αυτοάνοσης εγκεφαλομυελίτιδας (EAE) σε μοντέλα ποντικών (Bai et al., 2004; Liu et al., 2007b).

2.4.2 Μέθοδοι

Έχουν πραγματοποιηθεί πολλές μελέτες ασθενών - μαρτύρων για τη διερεύνηση της πιθανής συσχέτισης των πολυμορφισμών του γονιδίου CD24 με την ανάπτυξη και εξέλιξη της ΣκΠ, αλλά τα αποτελέσματα είναι αμφιλεγόμενα. Η παρούσα μελέτη έχει ως στόχο τη διεξαγωγή μιας μετα-ανάλυσης ώστε να δοθεί στατιστική ισχύς σε αυτές τις συσχετίσεις. Προς αυτή την κατεύθυνση, πραγματοποιήθηκε εκτεταμένη αναζήτηση βιβλιογραφίας, μέχρι τις 8-9-2014, στη βάση δεδομένων

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

PubMed με λέξεις κλειδιά «CD24 AND (GENE OR VARIANT OR polymorphism OR mutant OR mutation OR allele) AND ("Multiple sclerosis" OR MS OR "disseminated sclerosis" OR "encephalomyelitis disseminate")». Ως μέγεθος επίδρασης της μετα-ανάλυσης χρησιμοποιήθηκε ο σχετικός λόγος των συμπληρωματικών πιθανοτήτων (Odds Ratio, OR) και τα 95% διαστήματα εμπιστοσύνης, σύμφωνα με το μοντέλο τυχαίων επιδράσεων (DerSimonian and Laird, 1986). Συμπεριλήφθησαν και δεδομένα πυρηνικών οικογενειών (family trios) όπου και οι δυο γονείς και τουλάχιστον ένα παιδί πάσχουν από ΣκΠ σύμφωνα με πρόσφατη μέθοδο (Bagos, 2011). Πραγματοποιήθηκε επίσης πολυμεταβλητή μετα-ανάλυση για να αποδώσει ακριβέστερα το γενετικό μοντέλο κληρονομικότητας (Bagos, 2008). Για τη διεξαγωγή των παραπάνω στατιστικών μεθόδων χρησιμοποιήθηκε το στατιστικό πακέτο Stata.

2.4.3 Αποτελέσματα

Από την αναζήτηση της βιβλιογραφίας ανακτήθηκαν 19 άρθρα από τα οποία μόνο 7 άρθρα που περιείχαν 8 μελέτες εκπλήρωναν τα κριτήρια επιλογής και έδιναν δεδομένα για την πραγματοποίηση της παρούσας μετα-ανάλυσης. Ειδικότερα, οι οκτώ μελέτες χρησιμοποιήθηκαν για τη διερεύνηση του πολυμορφισμού CD24 226C>T (Ala57Val) και περιλάμβαναν συνολικά 2002 ασθενείς και 2185 υγιείς. Για τη διερεύνηση καθενός από τους πολυμορφισμούς 1527-1528 TG>del, 1056 A>G και 1626 A>G χρησιμοποιήθηκαν δυο άρθρα τα οποία περιείχαν 377 ασθενείς και 648 μάρτυρες.

Πίνακας 2.3. Οι πολυμορφισμοί του γονιδίου CD24 που διερευνήθηκαν για τη συσχέτιση με τη Σκλήρυνση κατά Πλάκας.

Πολυμορφισμοί	Αριθμός Μελετών	Ασθενείς	Υγιείς
Ala57Val, 226C>T	8	2002	2185
1527/1528 TG>del	2	377	648
1056 A>G	2	377	648
1626 A>G	2	377	648

Αρχικά, η μετα-ανάλυση για τον πολυμορφισμό 226C>T (Ala57Val) αποκάλυψε στατιστικά σημαντική συσχέτιση με την ΣκΠ με ORs πάντα μεγαλύτερα της μονάδας όπως φαίνεται και στον Πίνακα 2.4. Στη μετα-ανάλυση για τη σύγκριση αλληλόμορφων συμπεριλήφθησαν δεδομένα και από family trios. Στη μετα-ανάλυση για τη σύγκριση TT vs TC + CC συμπεριλήφθησαν δεδομένα από 8 μελέτες ενώ στη μετα-ανάλυση για τη σύγκριση TT + TC vs CC συμπεριλήφθησαν δεδομένα από 6

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

μόνο μελέτες, επειδή δύο μελέτες έδιναν στοιχεία μόνο για συνδυασμό γονότυπων και όχι για κάθε ένα ξεχωριστά. Τα αποτελέσματα ήταν προς την ίδια κατεύθυνση της σημαντικής συσχέτισης και όταν οι μετα-αναλύσεις έγιναν με δεδομένα από πληθυσμούς που βρίσκονταν σε ισορροπία Hardy-Weinberg. Η ετερογένεια ήταν μεγάλη για τις συγκρίσεις σύμφωνα με το συνεπικρατές και υπολειπόμενο μοντέλο κληρονομικότητας ενώ για το επικρατές η ετερογένεια ήταν πολύ χαμηλή (πίνακας 2.4).

Πίνακας 2.4. Μονομεταβλητή μετα-ανάλυση για τον πολυμορφισμό 226C>T (Ala57Val) του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας.

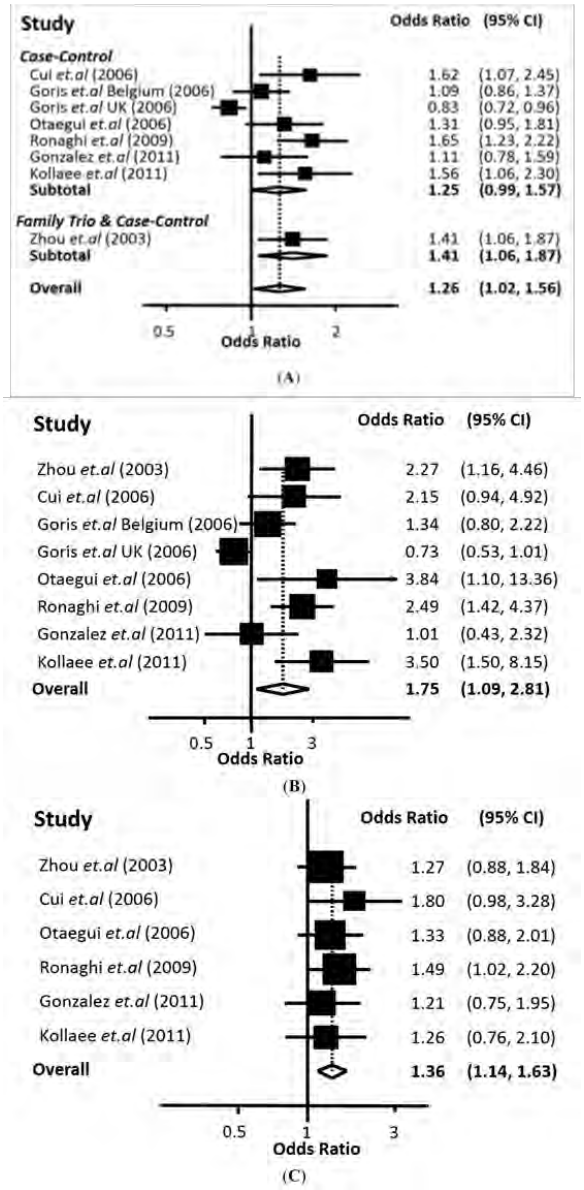
Συνδυασμός Γονότυπων (Contrast)	Μοντέλο Κληρονομικότητας	Αριθμός Μελετών	OR (μοντέλο τυχαίων επιδράσεων)	95% Διαστημα εμπιστοσύνης (CI)	I^2(%)	Cochran's Q	BSV	Z
T *vs* C	Συνεπικρατές	8	1.26	1.02 - 1.56	78.0%	31.87	0.07	2.16
T *vs* C in HWE	Συνεπικρατές	4	1.39	1.17 - 1.66	0.0%	2.33	0.00	3.70
TT *vs* TC+CC	Υπολειπόμενο	8	1.75	1.09 - 2.81	76.9%	30.25	0.33	2.34
TT *vs* TC+CC in HWE	Υπολειπόμενο	4	2.05	1.27 - 3.31	32.1%	4.42	0.08	2.93
TT+TC *vs* CC	Επικρατές	6	1.36	1.14 - 1.63	0.0%	1.50	0.00	3.34
TT+TC *vs* CC in HWE	Επικρατές	4	1.32	1.05 – 1.67	0.0%	1.21	0.00	2.33

Για να αποσαφηνιστεί ο τρόπος κληρονομικότητας πραγματοποιήθηκε πολυμεταβλητή μετα-ανάλυση με τις έξι μελέτες. Βρέθηκε ότι η σύγκριση TT vs CC δίνει στατιστικά σημαντική συσχέτιση με OR 2.44 και 95% CI 1.77 - 3.38 (Πίνακας 2.5). Επιπλέον, ο λόγος λογαρίθμων πιθανοτήτων λ=b1/b2 είναι 0.18, δηλαδή <1, το οποίο υποδηλώνει υπολειπόμενο τρόπο κληρονομικότητας. Επιπροσθέτως, ελέγχοντας τη μηδενική υπόθεση βρέθηκε ότι η συσχέτιση είναι πολύ ισχυρή με p-value <10-4. Έτσι, ο πολυμορφισμός 226C>T του γονιδίου CD24 συσχετίζεται με στατιστικά σημαντικό τρόπο με τη ΣκΠ με υπολειπόμενο τρόπο κληρονομικότητας.

Πίνακας 2.5. Πολυμεταβλητή μετα-αναλυση για τον πολυμορφισμό 226C>T (Ala57Val) του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας.

Συνδυασμός Γονότυπων (Contrast)	Αριθμός Μελετών	OR	95% CI
TC *vs* CC	6	1.18	0.97 - 1.42
TT *vs* CC	6	2.44	1.77 - 3.38

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 2.3. Τα forest plots της μετα-αναλύσης για τον πολυμορφισμό 226C>T (Ala57Val) του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας. Α) για το συνεπικρατές μοντέλο κληρονομικότητας Β) για το υπολειπόμενο C) για το επικρατές.

Στη συνέχεια, πραγματοποιήθηκε μετα-ανάλυση για τον πολυμορφισμό 1527-1528 TG>del στην οποία χρησιμοποιήθηκαν μόνο δυο μελέτες. Βρέθηκε συσχέτιση της διαγραφής (del) με την ΣκΠ εφόσον η σύγκριση del vs TG παρείχε OR 0.60 με 95% CI 0.41 - 0.85 και η σύγκριση del/del+TG/del vs TG/TG (επικρατής τρόπος κληρονομικότητας) έδωσε OR=0.60 με 95% CI 0.39 - 0.83. Σε αντίθεση, η σύγκριση del/del vs TG/TG+TG/del έδωσε μη στατιστικά σημαντική συσχέτιση (Πίνακας 2.6).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.6. Μονομεταβλητή μετα-αναλυση για τον πολυμορφισμό 1527-1528 TG>del του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας.

Συνδυασμός Γονότυπων (Contrast)	Μοντέλο Κληρονομικότητας	Αριθμός Μελετών	OR	95% CI	I^2(%)	Cochran's Q	BSV	z
del vs TG	Συνεπικρατές	2	0.60	0.41 - 0.80	0.0%	0.55	0.00	2.8
del/del *vs* TG/TG+TG/del	Υπολειπόμενο	2	1.09	0.06 – 18.64	55.2%	2.23	0.00	0.06
del/del +TG/del *vs* TG/TG	Επικρατές	2	0.57	0.39 -0.83	0.0%	0.2	0.00	2.89

Για την περαιτέρω διαλεύκανση της συσχέτισης, πραγματοποιήθηκε πολυμεταβλητή μετα-ανάλυση η οποία έδωσε στατική ισχύ στη σύγκριση del/TG vs TG/TG (OR 0.57 με 95% CI 0.38-0.87 (overall p-value <0.038). Από τα αποτελέσματα και των δυο μετα-αναλύσεων προκύπτει ότι η απαλοιφή του δινουκλεοτιδίου λειτουργεί προστατευτικά για την ανάπτυξη ΣκΠ, ενώ ο τρόπος κληρονόμησης φαίνεται να είναι ο επικρατής και υπάρχουν ενδείξεις για υπερεπικράτηση (λ=1.97>1) (Πίνακας 2.7).

Πίνακας 2.7. Πολυμεταβλητή μετα-ανάλυση για τον πολυμορφισμό 1527-1528 TG>del του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας.

Συνδυασμός Γονότυπων (Contrast)	Αριθμός Μελετών	OR	95% CI
del/TG *vs* TG/TG	2	0.57	0.38 – 0.87
del/del *vs* TG/TG	2	0.75	0.05 – 10.37

Τέλος έγιναν μετα-αναλύσεις δεδομένων από δυο μελέτες για άλλους δυο πολυμορφισμούς του γονιδίου CD24, τους 1056 A>G και 1626 A>G. Για κανέναν πολυμορφισμό και σε καμία σύγκριση δεν βρέθηκε στατιστικά σημαντική συσχέτιση με την ΣκΠ (Πίνακας 2.8).

2.4.4 Συμπεράσματα

Συμπερασματικά, τα αποτελέσματά μας δείχνουν σαφώς γενετική συσχέτιση μεταξύ του μη συνώνυμου πολυμορφισμού 226C>T του γονιδίου CD24 (Ala57Val) με τη ΣκΠ, ενώ προκύπτει ότι ο τρόπος κληρονομικότητας είναι ο υπολειπόμενος. Προτείνουμε ότι ο συγκεκριμένος πολυμορφισμός μπορεί να χρησιμοποιηθεί ως βιοδείκτης στην εκτίμηση κινδύνου για εμφάνιση ΣκΠ. Ο πολυμορφισμός 1527-

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

1528 TG>del φαίνεται να παίζει ένα προστατευτικό ρόλο κατά της ΣκΠ και μάλιστα μεταβιβάζεται με επικρατή τρόπο. Όσον αφορά τους πολυμορφισμούς 1056 A>G και 1626 A>G δεν φαίνεται να σχετίζονται με την ανάπτυξη της ΣκΠ. Ωστόσο, λόγω του μικρού μεγέθους του δείγματος τα αποτελέσματα για τους τρεις τελευταίους πολυμορφισμούς δεν θα πρέπει να υπερεκτιμούνται. Επιπλέον, πρόσθετες μελέτες για τη διερεύνηση της ανισορροπίας σύνδεσης μεταξύ των πολυμορφισμών του γονιδίου CD24 ή μελέτες αλληλεπιδράσεις γονιδίου-γονιδίου ή ακόμα και αλληλεπιδράσεων γονιδίων-περιβάλλοντος θα ήταν επωφελείς για την καλύτερη κατανόηση του ρόλου του γονιδίου CD24 στη εμφάνιση και την εξέλιξη της νόσου.

Πίνακας 2.8. Μονομεταβλητή μετα-ανάλυση για τους πολυμορφισμούς 1056 A>G and 1626 A>G του γονιδίου CD24 με τη Σκλήρυνση κατά Πλάκας.

Πολυμορφισμός	**Συνδυασμός Γονότυπων (Contrast)**	**Αριθμός Μελετών**	**OR**	**95% CI**	**I^2(%)**	**Cochran's Q**	**BSV**	**Z**
1056 A>G	G *vs* A	2	1.70	0.68 – 4.23	96.0%	24.96	0.42	1.13
	GG *vs* GA+AA	2	2.23	0.58 - 8.61	95.5%	22.26	0.91	1.16
	GG+GA *vs* AA	2	1.59	0.69 – 3.62	86.3%	7.32	0.31	1.09
1626 A>G	G *vs* A	2	0.79	0.61 – 1.02	0.0%	0.04	0.0	1.82
	GG *vs* GA+AA	2	0.61	0.24 - 1.58	0.0%	0.12	0.0	1.02
	GG+GA *vs* AA	2	0.78	0.58 – 1.04	0.0%	0.08	0.0	1.71

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

2.5 Μετα-ανάλυση των πολυμορφισμών του γονιδίου FZD3 με τον κίνδυνο εμφάνισης της σχιζοφρένειας

2.5.1 Εισαγωγή

Η σχιζοφρένεια είναι μια ψυχική διαταραχή που χαρακτηρίζεται από ένα πλήθος συμπτωμάτων όπως: παραισθήσεις, ψευδαισθήσεις, διαταραχές σκέψης, διαταραχές της κίνησης, διαταραχή της συμπεριφοράς, προβλήματα με διάσπαση προσοχής, προβλήματα με τη μνήμη, η κακή εκτελεστική λειτουργία. Παρά το γεγονός ότι η σχιζοφρένεια έχει προσελκύσει το επιστημονικό ενδιαφέρον για περισσότερο από έναν αιώνα, θεμελιώδη ερωτήματα όπως τι προκαλεί σχιζοφρένεια, αν η σχιζοφρένεια μπορεί να θεωρηθεί μια ασθένεια ή αν είναι μια κατάσταση ή περισσότερες, παραμένουν αναπάντητα (Bentall, 2013). Η εμφάνιση της σχιζοφρένειας προκαλείται από διάφορους παράγοντες (γονίδια, περιβαλλοντικοί παράγοντες, διαφορετική χημεία του εγκεφάλου και διαφορετική δομή του) οι οποίοι μπορούν να δρουν είτε μεμονωμένα είτε συνδυαστικά. Ωστόσο, αυτό που είναι γνωστό είναι ότι η σχιζοφρένεια είναι μια οικογενειακή ασθένεια (κληρονομική) (Lichtenstein et al., 2009; Sullivan et al., 2003). Δεδομένου ότι τα γονίδια κληρονομούνται μέσα στην οικογένεια, υπάρχει έντονο ενδιαφέρον για τη διερεύνηση της πιθανής συσχέτισης πολλών γενετικών πολυμορφισμών και της σχιζοφρένειας. Η ανθρώπινη πρωτεΐνη που κωδικοποιείται από το γονίδιο Frizzled Class Receptor 3 (FZD3) είναι ένας υποδοχέας επτά διαμεμβρανικών α-ελίκων και ανήκει στην οικογένεια Frizzled. Έχει μια μεγάλη εξωκυτταρική αμινοτελική περιοχή πλούσια σε κυστεΐνες (cysteine-rich domain, CRD). Τα μέλη της πρωτεϊνικής οικογένειας Frizzled είναι υποδοχείς για τις εκκρινόμενες WNT γλυκο-πρωτεΐνες (τύπου Wingless) που εμπλέκονται στον έλεγχο της ανάπτυξης και ανήκουν στην ευρύτερη οικογένεια υποδοχέων συζευγμένων με G-πρωτεΐνες (GPCRs) (ChuanYuan et al., 2011). Το γονίδιο FZD3, χαρτογραφείται στο χρωμόσωμα 8p21, αποτελείται από οκτώ εξώνια και είναι μήκους περίπου 70 kb (Jeong et al., 2006). Πολυμορφισμοί της περιοχής 8p21-22 έχουν συσχετιστεί με αυξημένο κίνδυνο εμφάνισης της σχιζοφρένειας (Pulver et al., 1995) κάτι που επαληθεύτηκε από μια πιο πρόσφατη μελέτη (Yang et al., 2003). Παρόμοιες μελέτες με αυτή του Yang και των συνεργατών του (Yang et al., 2003) ακολούθησαν τα επόμενα χρόνια και για άλλους πληθυσμούς, καταλήγοντας όμως σε αντικρουόμενα συμπεράσματα. Έχοντας ως στόχο τη διερεύνηση της σχέσης των πολυμορφισμών του γονιδίου FZD3 με τη σχιζοφρένια πραγματοποιήθηκε μια μετα-ανάλυση ώστε να συνδυαστούν τα διαθέσιμα δεδομένα των δημοσιευμένων γενετικών μελετών, οι οποίες πιθανόν μεμονωμένα να αποτυγχάνουν να δείξουν την πραγματική σχέση του εν λόγω γονιδίου με τη σχιζοφρένεια λόγω μειωμένης ισχύος.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

2.5.2 Μέθοδοι

Χρησιμοποιώντας ως λέξεις κλειδιά τους όρους "FZD3 ή Frizzled-3" και "Schizophrenia" πραγματοποιήθηκε αναζήτηση στις βιβλιογραφικές βάσεις δεδομένων Pubmed και Scopus. Στη μετα-ανάλυση συμπεριλήφθησαν μελέτες οι οποίες περιείχαν επαρκή στοιχεία για τον υπολογισμό του σχετικού λόγου συμπληρωματικών πιθανοτήτων (OR) του γονιδίου FZD3 με τη σχιζοφρένεια. Επίσης, συμπεριλήφθησαν μελέτες γενετικής συσχέτισης τόσο σε οικογένειες όσο και σε πληθυσμούς. Από κάθε μελέτη καταχωρήθηκε το όνομα του πρώτου συγγραφέα, το έτος δημοσίευσης της μελέτης, η καταγωγή του πληθυσμού και η χώρα, ο σχεδιασμός της μελέτης, ο αριθμός των family trios και ο αριθμός των ασθενών και υγιών μαζί με τα χαρακτηριστικά αυτών. Καθώς οι μελέτες που συμπεριλάβαμε είχαν διεξαχθεί κάτω από δυο διαφορετικούς επιδημιολογικούς σχεδιασμούς, -μελέτες ασθενών–μαρτύρων σε οικογένειες και σε πληθυσμούς-, υιοθετήθηκε η μονομεταβλητή μέθοδος μετα-ανάλυσης δεδομένων η οποία προτείνεται από τους (Bagos et al., 2011; Janssen et al., 2007). Η σύγκριση των αλληλόμορφων του γονιδίου σε ασθενείς και μάρτυρες έγινε με το σχετικό λόγο των συμπληρωματικών πιθανοτήτων (Odds Ratio, OR) μαζί με το 95% διάστημα εμπιστοσύνης αυτού (95% CI). Επιπλέον πραγματοποιήθηκε έλεγχος συστηματικού σφάλματος δημοσίευσης, διαχρονικής τάσης και υπολογίστηκε ο αριθμός των πρόσθετων ατόμων που θα έπρεπε να συμπεριληφθούν στη μελέτη ώστε τα αποτελέσματα της μετα-ανάλυσης να είναι στατιστικά σημαντικά. Η μετα-ανάλυση πραγματοποιήθηκε με το στατιστικό πακέτο Stata 13 και το επίπεδο σημαντικότητας ορίστηκε α=0.05.

2.5.3 Αποτελέσματα

Από την αναζήτηση της βιβλιογραφίας ανακτήθηκαν 15 μελέτες από τη βάση δεδομένων Pubmed και 4 μελέτες από τη Scopus. Από αυτές, οι 9 πληρούσαν τα κριτήρια για να συμπεριληφθούν στην ανάλυση και παρείχαν πληροφορίες για 6 μονονουκλεοτιδικούς πολυμορφισμούς του γονιδίου FZD3 (rs2241802, rs2323019, rs352203, rs3757884, rs880481 και rs960914). Οι μελέτες περιείχαν δεδομένα για 366 family trios, 1917 ασθενείς με σχιζοφρένεια και 2192 υγιείς. Όλες οι μελέτες αξιολογούσαν περισσότερους από έναν γονιδιακούς πολυμορφισμούς. Μια μελέτη περιείχε πληροφορίες και για τους δυο σχεδιασμούς (σε οικογένειες και σε πληθυσμούς) (Ide et al., 2004), δυο μελέτες περιελάμβαναν μόνο family trios (Wei and Hemmings, 2004; Yang et al., 2003), ενώ οι υπόλοιπες αφορούσαν μελέτες ασθενών- μαρτύρων σε πληθυσμούς. Στον Πίνακα 2.9 παρουσιάζονται τα χαρακτηριστικά των μελετών που συμμετείχαν στη μετα-ανάλυση για κάθε πολυμορφισμό του γονιδίου FZD3.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.9. Χαρακτηριστικά των μελετών που περιλαμβάνονται στην μετα-ανάλυση των πολυμορφισμών του γονιδίου FZD3και τη συσχέτιση τους με τη σχιζοφρένεια.

Study	Year	Case characteristics	Controls characteristics	Family trios	Descent	SNP markers
Yang (Yang et al., 2003)	2003			246 family trios consisting (patients: 138 (56%) men and 108 (44%) women). Patients met ICD-10 criteria. Mean duration of illness: 5 years. Allelic association for individual SNPs was accessed by transmission disequilibrium test (TDT)	Chinese, Han	rs181277, rs2241802, rs2323019, rs2874940, rs352203
Katsu (Katsu et al., 2003)	2003	200 unrelated patients (108 men and 92 women; mean age 45.1 ± 12.5 years) fulfilling the ICD-10 criteria, 91 were diagnosed with paranoid type, 107 with hebephrenic type and 2 with catatonic type schizophrenia	218 healthy, gender, age, and geographical origin matched volunteers (112 men and 106 women; mean age 46.5 ± 15.7 years), mostly from the medical staff, subjects with a positive personal or familial history of major psychiatric disorders were excluded		Japanese	rs2241802, rs3757888, rs960914
Ide (Ide et al., 2004)	2004	540 unrelated patients (270 men, 270 women; mean age, 45.6 ± 11.0 years). Consensus diagnosis was made according to the DSM-IV criteria	540 age and gender-matched control subjects with no history of mental illness (270 men, 270 women; mean age, 45.1 ±11.7 years)	212 schizophrenia families (643 members), including 168 independent and complete trios	Japanese	rs2241802, rs2323019, rs352203, rs352210, rs352222, rs352226, rs3757884, rs960914
Zhang (Zhang et al., 2004)	2004	236 patients with schizophrenia (130 men and 106 women; mean age: 27±7.64 years) diagnosed according to ICD-10 criteria	275 healthy controls (158 men and 117 women; mean age: 28.34±8.95 years) matched to cases for age, sex, and ethnicity, with no history of mental health and neurological diseases		Chinese, Han	rs2241802, rs2323019, rs352203, rs880481
Wei (Wei and Hemmings, 2004)	2004			120 family trios (82 males and 38 females), aged 29.03 ± 6.95 years, diagnosed by the psychiatrists who treated them during the period 1990 – 2002, allelic association for individual SNPs was accessed by transmission disequilibrium	English, Welsh, Irish and Scottish	rs2241802, rs2323019, rs352203

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

			test (TDT)		
Hashimoto (Hashimoto et al., 2005)	2005	427 patients (221 males and 206 females with mean age of 44.2± 14.5 years), 91 with bipolar disorder and 396 with major depression Consensus diagnosis was made according to the DSM-IV criteria	473 healthy controls (228 and 245; 36.1 ±12.5 years) with no current or past contact to psychiatric services	Japanese	rs2241802, rs2323019, rs352203, rs960914
Jeong (Jeong et al., 2006)	2006	241 unrelated patients (147 males and 94 females, average age=32.3±8.38) diagnosed with schizophrenia by attending psychiatrists	192 healthy subjects (125 males and 67 females, average age=22.6±6.14) who were hospital staff and college student volunteers. Subjects who had first degree relatives with suspected psychiatric illness were excluded	Korean	rs2241802, rs960914
Reif (Reif et al., 2007)	2007	192 unrelated patients diagnosed according to ICD-10 criteria and Leonhard's system	284 healthy blood donors coming from the same recruiting area as the patients	German	rs2241802, rs352203, rs960914
ChuanYuan (ChuanYuan et al., 2011)	2011	81 patients (52% male and 48% female, mean age=33.8±10.9 years) meting the DSM-IV 42 criteria	210 healthy subjects (107 male and 103 female, mean age 34.10 ±10.33 years) matched for age and gender to patients, with no any biological relationship with patients	Chinese, Va	rs2241802, rs2323019, rs352203, rs3757888, rs880481

Ο πολυμορφισμός του γονιδίου rs2241802 εξετάστηκε για το σύνολο των 9 μελετών που περιλαμβάνονται στη μετα-ανάλυση. Δυο μελέτες είχαν πραγματοποιηθεί σε Ευρωπαίους ενώ οι υπόλοιπες σε Ασιάτες. Η σύγκριση του Α έναντι του C αλληλομόρφου έδωσε ένα μη στατιστικά σημαντικό Odds Ratio (0.90, 95% C.I.: 0.79,1.03). Οι έλεγχοι του Begg και του Egger δεν έδειξαν την ύπαρξη συστηματικού σφάλματος δημοσίευσης. Ο έλεγχος της αθροιστικής μετα-ανάλυσης παρείχε ενδείξεις διαχρονικής τάσης (p-value<0.001). Ο εκτιμώμενος αριθμός των επιπρόσθετων ατόμων που απαιτούνται για να προκύψουν στατιστικά σημαντικές εκτιμήσεις ήταν αρκετά μεγάλος (3228 άτομα) γεγονός που δηλώνει ότι είναι αρκετά δύσκολο να αλλάξει το αποτέλεσμα της μετα-ανάλυσης.

Στη μετα-ανάλυση του πολυμορφισμού rs2323019 συμπεριλήφθησαν 6 μελέτες που αξιολογούσαν την πιθανή επίδρασή του στην εμφάνιση σχιζοφρένειας με ένα σύνολο 366 family trios (2 μελέτες σε οικογένειες), 1284 ασθενείς με σχιζοφρένεια και 1498 υγιείς (4 μελέτες ασθενών-μαρτύρων σε πληθυσμούς). Η μετα-ανάλυση για τη σύγκριση του αλληλομόρφου Α έναντι του G κατέδειξε μια

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

οριακά μη σημαντική συσχέτιση με Odds Ratio ίσο με 1.21 (95% C.I.: 0.98, 1.50).). Οι έλεγχοι του Begg και του Egger δεν έδειξαν την ύπαρξη συστηματικού σφάλματος δημοσίευσης. Ο έλεγχος της αθροιστικής μετα-ανάλυσης παρείχε ενδείξεις διαχρονικής τάσης (p-value<0.001). Η εφαρμογή της μεθοδολογίας που παρουσιάστηκε από τον Barrowman και τους συνεργάτες του (Barrowman, Fang et al. 2003) έδειξε ότι απαιτούνται 772 επιπλέον άτομα για να βρεθεί στατιστικά σημαντική συσχέτιση του πολυμορφισμού με την ασθένεια.

Επτά μελέτες χρησιμοποιήθηκαν στη μετα-ανάλυση του πολυμορφισμού rs352203 με την σχιζοφρένεια περιλαμβάνοντας 366 family trios (2 μελέτες), 1476 σχιζοφρενείς και 1782 υγιείς (5 μελέτες). Από τη σύγκριση του αλληλομόρφου C έναντι του T εκτιμήθηκε ένα Odds Ratio ίσο με 0.82 (95% C.I.: 0.69, 0.98) υποδηλώνοντας ένα προστατευτικό ρόλο του υπό μελέτη πολυμορφισμού με τη νόσο. Οι έλεγχοι του Begg και του Egger δεν έδειξαν την ύπαρξη συστηματικού σφάλματος δημοσίευσης. Με βάση τα ευρήματα της αθροιστικής μετα-ανάλυσης, το Odds Ratio της πρώτης μελέτης (Yang, Si et al. 2003) διαφέρει σημαντικά από τις υπόλοιπες μελέτες.

Ο πολυμορφισμός rs3757888 εξετάστηκε σε δυο μόνο μελέτες (Katsu, Ujike et al.2003; ChuanYuan, Li et al. 2011) οι οποίες περιείχαν μόνο ασθενείς-μαρτύρες σε Ασιάτικο πληθυσμό. Συνολικά, η μετα-ανάλυση περιελάμβανε 709 άτομα, με 281 σχιζοφρενείς και 428 υγιείς. Το Odds Ratio της σύγκρισης του αλληλομόρφου G έναντι του A ήταν μη στατιστικά σημαντικό (OR: 1.08, 95% C.I.: 0.63, 1.84). Οι στατιστικοί έλεγχοι για το συστηματικό σφάλμα δημοσίευσης και για διαχρονική τάση δεν πραγματοποιήθηκαν δεδομένου του μικρού αριθμού μελετών. Τέλος, χρησιμοποιώντας τη μέθοδο του Barrowman και συνεργατών απαιτείται ένας εξαιρετικά μεγάλος αριθμός ατόμων για να επιτύχουμε στατιστικά σημαντικά αποτελέσματα (34032 άτομα).

Η επίδραση του πολυμορφισμού rs880481 στη σχιζοφρένεια περιελάμβανε 2 μελέτες ασθενών-μαρτύρων σε πληθυσμούς μόνο ασιατικής καταγωγής, -317 σχιζοφρενείς και 485 υγιείς. Από τη μετα-ανάλυση εκτιμήθηκε ένα μη σημαντικό Odds Ratio (OR: 1.22, 95% C.I.: 0.78, 1.91) για τη σύγκριση του αλληλομόρφου A έναντι του G με I2 ίσο με 71.6% (p-value=0.041). Ομοίως με τον πολυμορφισμό rs3757888 και εδώ δεν πραγματοποιήσαμε στατιστικούς ελέγχους για συστηματικό σφάλμα δημοσίευσης και διαχρονική τάση. Ο αριθμός των επιπλέον ατόμων για να υπολογιστούν στατιστικά σημαντικές επιδράσεις είναι 3364.

Ένα σύνολο 5 μελετών εξέταζαν τη συσχέτιση του πολυμορφισμού rs960914 με τον κίνδυνο εμφάνισης σχιζοφρένειας, από τις οποίες οι τέσσερις αφορούσαν Ασιάτες και μια Ευρωπαίους (Reif, Melchers et al. 2007). Η σύγκριση του αλληλομόρφου C έναντι του T οδήγησε σε ένα μη σημαντικό

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Odds Ratio (OR: 1.10, 95% C.I.: 0.93, 1.30). Οι έλεγχοι του Begg και του Egger δεν έδειξαν την ύπαρξη συστηματικού σφάλματος δημοσίευσης. Η μέθοδος των Barrowman και συνεργατών (Barrowman, Fang et al. 2003) έδειξε ότι απαιτούνται 7585 άτομα για στατιστικά σημαντικές εκτιμήσεις.

Τα αποτελέσματα δείχνουν μια δυνητικά ασθενέστερη επίδραση των πολυμορφισμών FZD3 στη σχιζοφρένεια σε σχέση με αυτή που είχε αρχικά προταθεί και, ενδεχομένως, να περιορίζεται σε πληθυσμούς Κινέζων. Δεν εντοπίστηκε συσχέτιση μεταξύ των πολυμορφισμών του γονιδίου FZD3 που εξετάστηκαν με τη σχιζοφρένεια, εκτός από τον πολυμορφισμό rs352203 (Εικόνα 2.4). Η σύγκριση του C αλληλόμορφου έναντι του T, έδειξε προστατευτική δράση για το υπολειπόμενο αλληλόμορφο C (OR=0.819, 95% CI: 0.69, 0.98) (Πίνακας 2.10). Ωστόσο, το αποτέλεσμα αυτό οφείλεται κυρίως σε μελέτες που περιελάμβαναν πληθυσμούς Κινέζων. Η ετερογένεια μεταξύ των μελετών για τους έξι πολυμορφισμούς που εξετάστηκαν στην παρούσα ανάλυση, όπως εκτιμάται με τη χρήση του δείκτη I^2, κυμάνθηκε από 54.4% έως 76.1%.

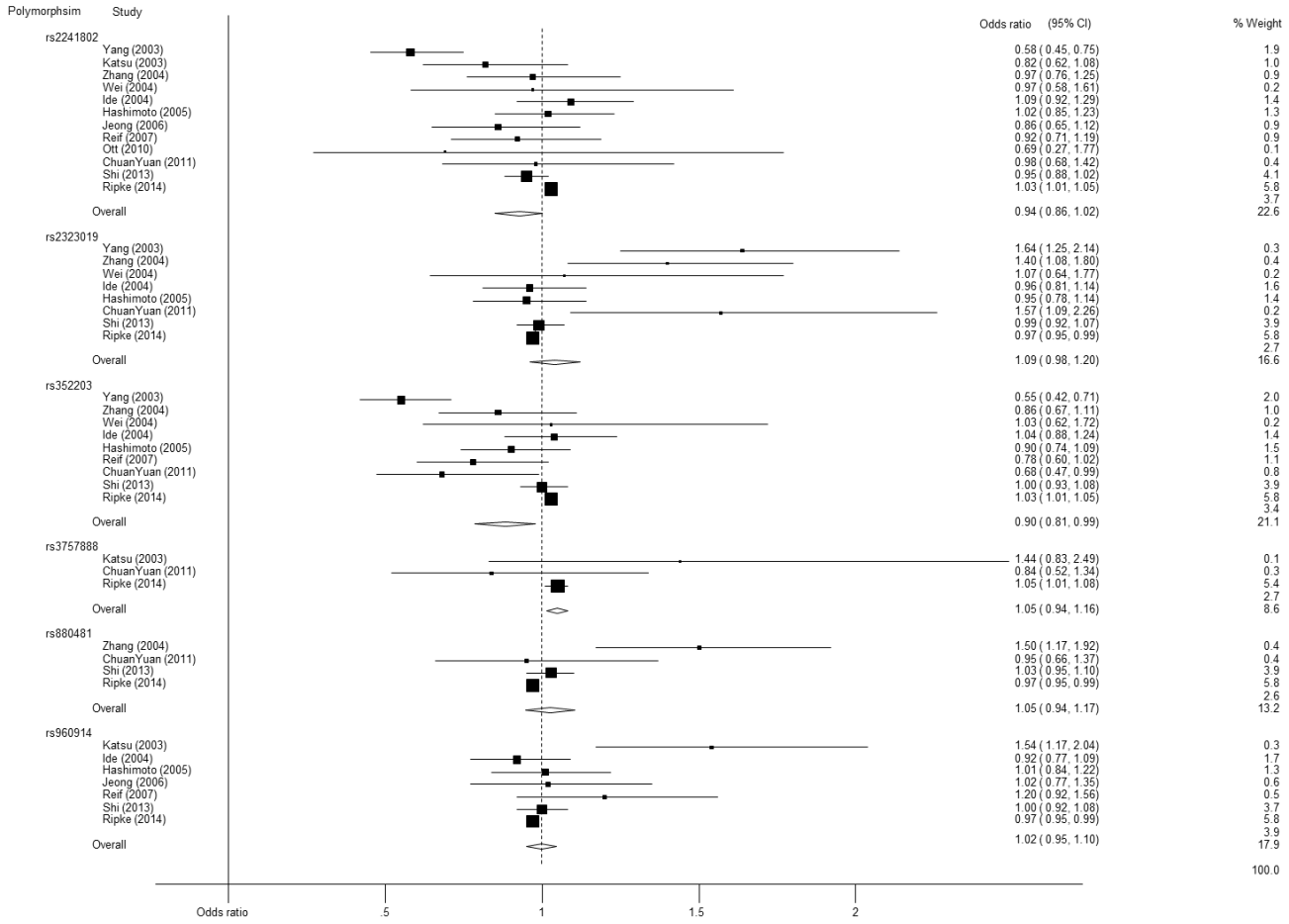
Πίνακας 2.10. Αποτελέσματα μετα-ανάλυσης των μελετών που εκτιμούν τη σχέση έξι πολυμορφισμών του γονιδίου FZD3 με τη σχιζοφρένεια.

Πολυμορφισμός	**Αριθμός μελετών**	**OR μοντέλου τυχαίων επιδράσεων (p-value)**	**95% Διάστημα Εμπιστοσύνης**	**I^2 (%) (p-value)**
rs2241802	9	0.90 (0.125)	0.79, 1.03	57.3 (0.022)
rs2323019	6	1.215 (0.073)	0.98, 1.50	75.0 (0.001)
rs352203	7	0.819 (0.024)	0.69, 0.98	67.7 (0.006)
rs3757888	2	1.08 (0.783)	0.63, 1.84	54.5 (0.183)
rs880481	2	1.217 (0.390)	0.78, 1.91	76.1 (0.041)
rs960914	5	1.10 (0.281)	0.93, 1.30	62.9 (0.029)

Σε πληθυσμούς Κινέζων εντοπίστηκε επίσης αυξημένος κίνδυνος για σχιζοφρένεια και στην περίπτωση του πολυμορφισμού rs2323019 (OR=1.52 , 95% CI: 1.29, 1.79, χωρίς ετερογένεια μεταξύ των μελετών) (Εικόνα 2.5). Δεδομένου ότι σε πληθυσμούς Κινέζων παρατηρείται διαφορετικό πρότυπο ανισορροπίας σύνδεσης σε σχέση με τους υπόλοιπους πληθυσμούς, είναι πιθανό, η ασθένεια της σχιζοφρένειας να σχετίζεται με άλλους πολυμορφισμούς του γονιδίου FZD3 που βρίσκονται σε

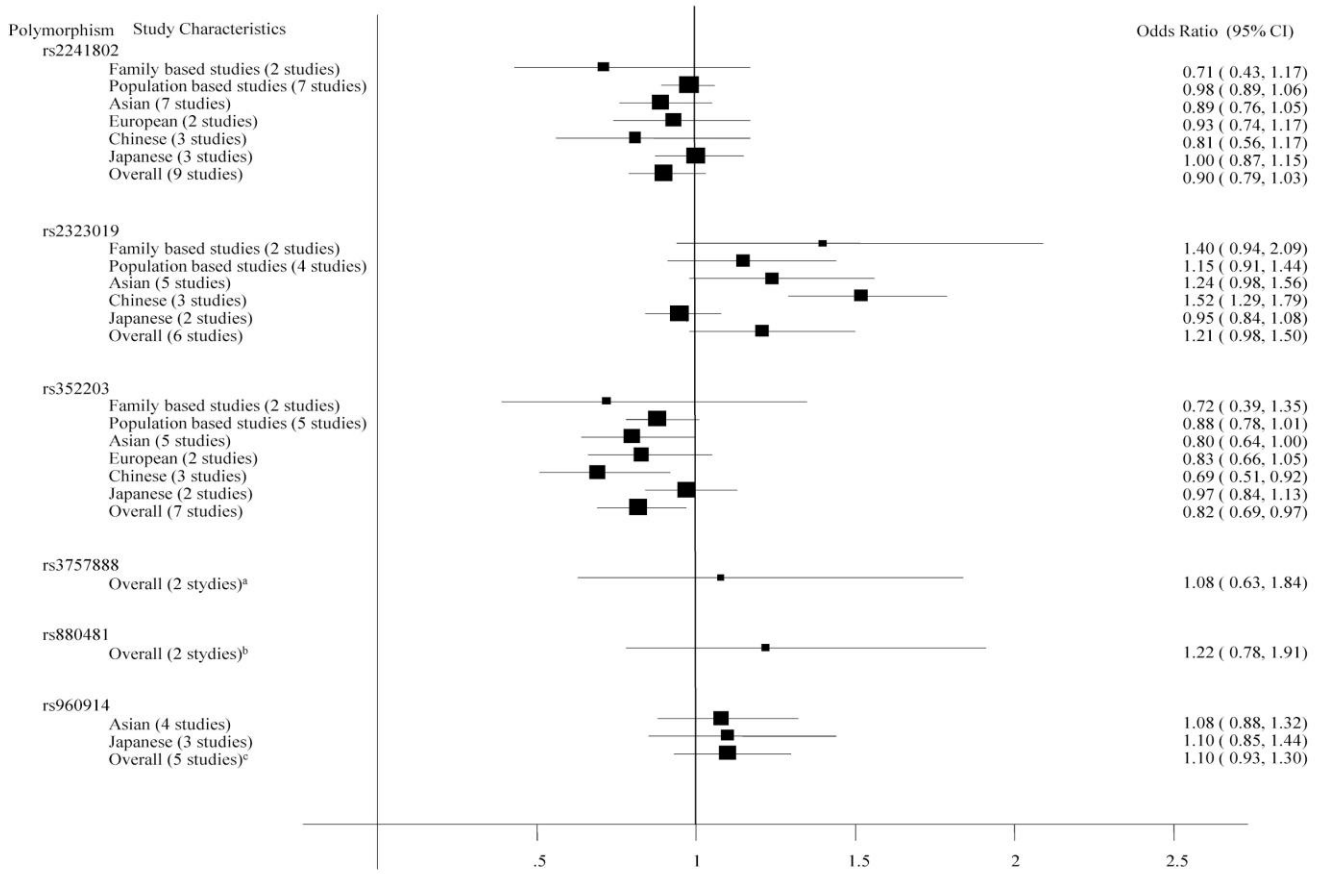
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ισχυρότερη ανισορροπία σύνδεσης στους Κινέζους από ό,τι στους υπόλοιπους πληθυσμούς που μελετήθηκαν.



Εικόνα 2.4. Γραφική απεικόνιση των αποτελεσμάτων της μετα-ανάλυσης για τη συσχέτιση των πολυμορφισμών rs2241802, rs2323019, rs352203, rs3757888, rs880481 και rs960914 του γονιδίου FZD3 με τη σχιζοφρένεια.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 2.5. Γραφική απεικόνιση των αποτελεσμάτων των μετα-αναλύσεων σε υποομάδες για τη συσχέτιση των πολυμορφισμών rs2241802, rs2323019, rs352203, rs3757888, rs880481 και rs960914 του γονιδίου FZD3 με την εμφάνιση σχιζοφρένειας. a και οι δυο μελέτες ήταν μελέτες ασθενών-μαρτύρων σε πληθυσμούς και περιελάμβαναν Ασιάτες; b και οι δυο μελέτες ήταν μελέτες ασθενών-μαρτύρων σε πληθυσμούς και περιελάμβαναν Κινέζους; c όλες οι μελέτες ήταν μελέτες ασθενών-μαρτύρων σε πληθυσμούς.

2.5.4 Συμπεράσματα

Η παρούσα μετα-ανάλυση, αν και απέτυχε να δείξει στατιστικά σημαντική συσχέτιση των πολυμορφισμών rs2241802, rs3757884, rs880481 και rs960914 με τη σχιζοφρένεια, υπογράμμισε το ρόλο του πολυμορφισμού rs352203. Πιο συγκεκριμένα η παρουσία του αλληλομόρφου C διαδραματίζει; έχει προστατευτικό ρόλο για την εμφάνιση της νόσου. Εκτιμήθηκε επίσης ένας αυξημένος κίνδυνος για τους φορείς του αλληλομόρφου A του πολυμορφισμού rs2323019 στους Κινέζους, κάτι που μπορεί να αποδοθεί στο διαφορετικό βαθμό ανισορροπίας σύνδεσης που εκτιμήθηκε για τη φυλή αυτή. Οι περιορισμοί της παρούσας μετα-ανάλυσης εστιάζονται κυρίως στο μικρό αριθμό μελετών με την πλειοψηφία αυτών να περιλαμβάνουν Ασιάτες. Τα αποτελέσματά μας

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

είναι προς την ίδια κατεύθυνση με εκείνα του Jeong και συνεργατών (Jeong et al., 2006) αν και μόνο δυο πολυμορφισμοί (rs2241802 και rs960914) είχαν μελετηθεί στην εργασία αυτή. Ωστόσο, εμείς επικαιροποιήσαμε την ανάλυση συμπεριλαμβάνοντας μεγαλύτερο αριθμό μελετών, υπολογίζοντας παράλληλα πιο ακριβείς εκτιμήσεις εφαρμόζοντας σύγχρονη μεθοδολογία που συνυπολογίζει τα αποτελέσματα από μελέτες ασθενών μαρτύρων σε πληθυσμούς και σε οικογένειες. Τέλος, υπολογίσαμε τα επιπρόσθετα άτομα που θα οδηγούσαν σε στατιστικά σημαντικές εκτιμήσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

2.6 Διερευνήση της γενετικής συσχέτισης των πολυμορφισμών των υποδοχέων της αγγειοτενσίνης με νεφρικές ασθένειες

2.6.1 Εισαγωγή

Η Χρόνια Νεφρική Νόσος (Chronic Kidney Disease) είναι το τελικό αποτέλεσμα πολλών νεφρικών διαταραχών, που χαρακτηρίζεται από μια αργή, προοδευτική και μη αναστρέψιμη επιδείνωση της νεφρικής λειτουργίας για πολλούς μήνες ή χρόνια αλλά συνήθως είναι ασυμπτωματική, δηλαδή ο ασθενής να διατηρεί καλή σωματική υγεία. Όταν ο ασθενής φτάσει στο τελικό στάδιο της ασθένειας (Νεφρική νόσος τελικού σταδίου) εκδηλώνεται μια χρόνια ουραιμία και είναι απαραίτητη η μεταμόσχευση ή η αιμοκάθαρση των νεφρών. Η νεφροπάθεια IgA (IgAN) και η κυστεοουρητηρική παλινδρόμηση (Vesicoureteral Reflux) είναι ασθένειες που μπορούν να οδηγήσουν σε Χρόνια Νεφρική Νόσο. Το σύστημα ρενίνης-αγγειοτενσίνης (RAS) παίζει έναν κρίσιμο ρόλο στην ρύθμιση των φυσιολογικών λειτουργιών του καρδιαγγειακού συστήματος. Το πρωταρχικό μόριο τελεστής αυτού του συστήματος, η αγγειοτενσίνη ΙΙ (ANGII), μέσω των υποδοχέων της (AGTR1 και AGTR2) επηρεάζει τη λειτουργία πολλών ζωτικών οργάνων συμπεριλαμβανομένων της καρδιάς, των νεφρών, των αγγείων και του εγκεφάλου (Mehta and Griendling, 2007). Οι πρωτεΐνες RAS είναι ρυθμιστές της αρτηριακής πίεσης και της ενδονεφρικής αιμοδυναμικής (Gumprecht et al., 2000). Η Αρτηριακή υπέρταση συχνά έχει βρεθεί να συσχετίζεται με τη χρόνια νεφρική νόσο και είναι ο πιο σημαντικός παράγοντας κινδύνου για την εξέλιξης της νεφρικής ανεπάρκειας (Basset el et al., 2002). Αν και ακόμα δεν έχουν αποσαφηνιστεί πλήρως οι παράγοντες που συμμετέχουν στη νεφρική ανεπάρκεια, υπάρχουν σαφείς ενδείξεις σχετικά με γενετικούς παράγοντες οι οποίοι εμπλέκονται άμεσα στην εμφάνιση της (McKnight et al., 2010). Στην παρούσα μελέτη, διερευνήθηκε η γενετική συσχέτιση των πολυμορφισμών των υποδοχέων της αγγειοτενσίνης με νεφρικές ασθένειες και πραγματοποιήθηκαν οι κατάλληλοι έλεγχοι ώστε να βρεθεί αν αυτοί οι πολυμορφισμοί μπορούν να χρησιμοποιηθούν ως προγνωστικοί δείκτες για τη νεφρική ανεπάρκεια.

2.6.2 Μέθοδοι

Πραγματοποιήθηκε μια ολοκληρωμένη αναζήτηση στη βιβλιογραφία μέχρι το Νοέμβριο του 2012 με τις λέξεις-κλειδιά: AGTR, AGTR1, AGTR1B, AGTR2, ‘ANGIOTENSIN RECEPTOR’, ‘ANGIOTENSIN II RECEPTOR’, GENE, VARIANT, POLYMORPHISM, MUTANT, MUTATION, ALLELE, ‘CHRONIC KIDNEY DISEASE’, ‘KIDNEY FAILURE’ ‘END-STAGE KIDNEY DISEASE’, ‘END-

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

STAGE RENAL DISEASE', 'END-STAGE RENAL FAILURE', DIALYSIS, 'IgA GLOMERULONEPHRITIS', 'IgA NEPHROPATHY', 'VESICOURETERAL REFLUX', VUR και με συνδυασμούς αυτών. Από κάθε μελέτη καταγράφηκε: ο κωδικός στη βάση δεδομένων Pubmed, το όνομα του πρώτου συγγραφέα, το έτος δημοσίευσης, η γεωγραφική θέση και η εθνικότητα του πληθυσμού που μελετήθηκε και ο συνολικός αριθμός των συμμετεχόντων (ασθενείς και υγιείς).

Οι κατανομές των αλληλομόρφων και των γονοτύπων υπολογίστηκαν στους ασθενείς και στους υγιείς σε κάθε μελέτη. Τα δεδομένα των επιμέρους μελετών συνδυάστηκαν χρησιμοποιώντας το μοντέλο τυχαίων επιδράσεων (DerSimonian and Laird, 1986), υπολογίστηκε ο σχετικός λόγος συμπληρωματικών πιθανοτήτων (odds Ratio/OR) με τα αντίστοιχα 95% διαστήματα εμπιστοσύνης, για κάθε γονότυπο ή αλληλόμορφο. Η ετερογένεια μεταξύ των μελετών αξιολογήθηκε με τη χρήση του ελέγχου χ^2 βασισμένο στον έλεγχο Q του Cochran και ο δείκτης ετερογένειας I^2 (Higgins et al., 2003). Επιπλέον, εφαρμόστηκε μια πολυπαραγοντική μέθοδος τυχαίων επιδράσεων (Bagos, 2008) με στόχο την εύρεση ισχυρών συσχετίσεων μεταξύ γονιδίων και ασθενειών. Για την εκτίμηση πιθανού συστηματικού σφάλματος δημοσίευσης χρησιμοποιήθηκαν οι έλεγχοι των Begg και Egger (Begg and Mazumdar, 1994; Egger et al., 1997). Τέλος, πραγματοποιήθηκε αθροιστική μετα-ανάλυση προκειμένου να διερευνηθεί αν υπάρχει το φαινόμενο της διαχρονικής τάσης και το φαινόμενο του Πρωτέα, κάτι το οποίο θα εισήγαγε συστηματικό σφάλμα (bias) στο τελικό αποτέλεσμα της μελέτης μας (Bagos and Nikolopoulos, 2009; Ioannidis and Trikalinos, 2005; Lau et al., 1992; Lau et al., 1995).

2.6.3 Αποτελέσματα

Η αναζήτηση στη βιβλιογραφία έγινε με σκοπό τον εντοπισμό όλων των μελετών που ερευνούν τη συσχέτισης των πολυμορφισμών των γονιδίων AGTR1 και το AGTR2 με νεφρικές παθήσεις. Μετα-ανάλυση πραγματοποιήθηκε για τους πολυμορφισμούς τους οποίου βρέθηκαν δεδομένα σε τουλάχιστον από τρεις μελέτες. Πιο συγκεκριμένα, για τη διερεύνηση της συσχέτισης του πολυμορφισμού A1166C του γονιδίου AGTR1 συμπεριλήφθησαν στη μετα-ανάλυση: 8 μελέτες για τη χρόνια νεφρική νόσο με 4252 υγιείς και 812 ασθενείς, 17 μελέτες για τη νεφρική νόσο τελικού σταδίου με 3866 υγιείς και 2596 ασθενείς, 5 μελέτες για τη νεφροπάθεια IgA με 1373 υγιείς και 785 ασθενείς και 3 μελέτες για την κυστεοουρητηρική παλινδρόμηση με 216 υγιείς και 174 ασθενείς. Για τον έλεγχο της συσχέτισης του πολυμορφισμού A1332G του γονιδίου AGTR2 με την κυστεοουρητηρική παλινδρόμηση χρησιμοποιήθηκαν 3 μελέτες με 790 υγιείς και 654 ασθενείς.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.11. Τα χαρακτηριστικά των μελετών που ανακτήθηκαν από την αναζήτηση στη βιβλιογραφία ανά γονίδιο, πολυμορφισμό και κατηγορία ασθένειας.

Ασθένεια	**Γονίδιο**	**Πολυμορφισμός**	**Ασθενείς / Υγιείς**	**Αριθμός Μελετών**
Νεφρική νόσος Τελικού Σταδίου	AGTR1	A1166C/rs5186	2596 / 3866	17
Νεφρική νόσος Τελικού Σταδίου	AGTR1	C521T		1
Νεφρική νόσος Τελικού Σταδίου	AGTR1	A1138T		1
Νεφρική νόσος Τελικού Σταδίου	AGTR1	AG214CC		1
Χρόνια Νεφρική Νόσος	AGTR1	A1166C/rs5186	812 / 4252	8
Χρόνια Νεφρική Νόσος	AGTR1	C573T		1
Χρόνια Νεφρική Νόσος	AGTR1	C521T		2
Χρόνια Νεφρική Νόσος	AGTR1	A1138T		1
Χρόνια Νεφρική Νόσος	AGTR1	AG214CC		1
Χρόνια Νεφρική Νόσος	AGTR1	G163A		2
Χρόνια Νεφρική Νόσος	AGTR2	A1332G/rs5194		1
Νεφροπάθεια IgA	AGTR1	A1166C/rs5186	785 / 1373	5
Κυστεοουρητηρική Παλινδρόμηση	AGTR1	A1166C/rs5186	174 / 216	3
Κυστεοουρητηρική Παλινδρόμηση	AGTR2	A1332G/rs5194	654 / 790	3

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Στη μετα-ανάλυση για τον έλεγχο της συσχέτισης του πολυμορφισμού A1166C του γονιδίου AGTR1 με τη νεφρική νόσο τελικού σταδίου δε βρέθηκε στατιστικά σημαντική συσχέτιση [OR 1,12 (95% CI: 0,90 - 1,38)] για τα αλληλόμορφα. Ομοίως, μη στατιστικά σημαντική συσχέτιση αποκαλύφθηκε και στους ελέγχους για το επικρατές και το υπολειπόμενο μοντέλο κληρονομικότητας. Επίσης, δε βρέθηκε συσχέτιση του πολυμορφισμού A1166C του γονίδιο AGTR1 με τη χρόνια νεφρική νόσο στη μετα-ανάλυση με τα αλληλόμορφα αλλά και σε αυτή με τους γονότυπους [OR 1,16 (95% CI: 0,83 - 1,64) για τα αλληλόμορφα]. Πρόσθετες μετα-αναλύσεις πραγματοποιήθηκαν για την εύρεση συσχέτισης του πολυμορφισμού A1166C του γονιδίου AGTR1 με τη Νεφροπάθεια IgA και την Κυστεοουρητηρική Παλινδρόμηση. Ελέγχοντας όλα τα πιθανά μοντέλα κληρονομικότητας δε βρέθηκε κάποια στατιστικά σημαντική συσχέτιση. Επιπλέον, πραγματοποιήθηκε πολυπαραγοντική ανάλυση και επιβεβαιώθηκε ότι δεν υπάρχει συσχέτιση των παραπάνω ασθενειών με τον πολυμορφισμό A1166C του γονιδίου AGTR1.

Ο πολυμορφισμός A1332G του γονιδίου AGTR2 εξετάστηκε για τη συσχέτιση του με την Κυστεοουρητηρική Παλινδρόμηση. Ωστόσο δε βρέθηκε στατιστικά σημαντική συσχέτιση μιας και το OR ήταν 1,02 (95% CI: 0.67- 1,55).

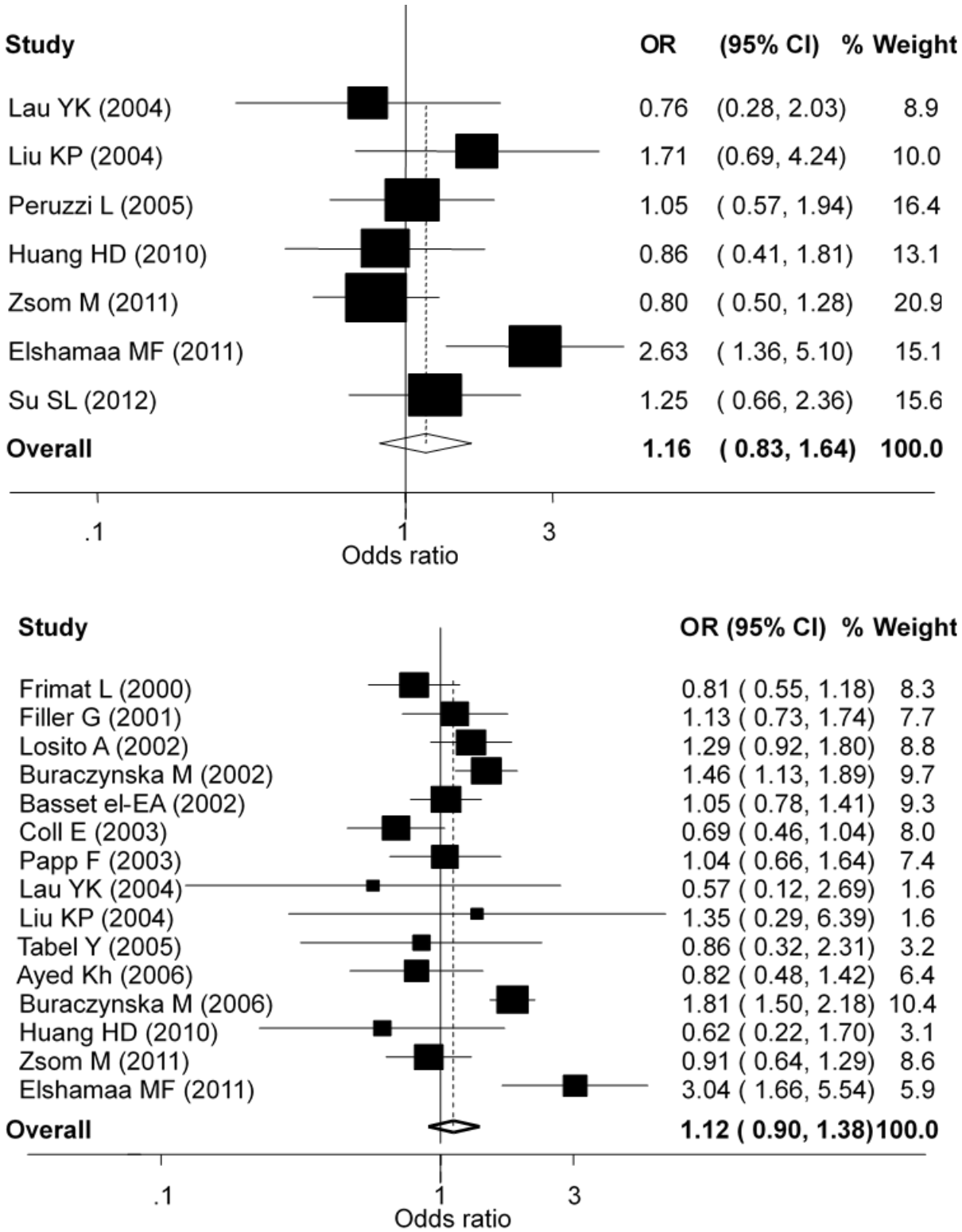
Επιπλέον, σε όλους του ελέγχους που πραγματοποιήθηκαν για όλες τις ασθένειες και για τους δυο πολυμορφισμούς δε βρέθηκε συστηματικό σφάλμα δημοσίευσης μεταξύ των μελετών. Τέλος, διερευνήθηκε αν η ύπαρξη του πολυμορφισμού A1166C του γονιδίου AGTR1 σχετίζεται με την εμφάνιση της υπέρτασης σε ασθενείς με χρόνια νεφρική νόσο και νεφρική νόσο τελικού σταδίου, αλλά καμία στατιστικά σημαντική συσχέτιση δε βρέθηκε, επίσης.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.12. Τα αποτελέσματα της μονοπαραγοντικής μετα-ανάλυση για όλα τα αλληλόμορφα των πολυμορφισμών των γονιδίων AGTR1και AGTR2 για τη συσχέτιση τους με νεφρικές παθήσεις.

	Πολυμορφισμός	**Ασθένεια**	**Αριθμός Μελετών**	**Odds Ratio**	**95% Διάστημα Εμπιστοσύνης**	**Cochran's Q**	**P-value Ετερογένειας**
A *vs* C	A1166C/AGTR1	Νεφρική νόσος Τελικού Σταδίου	16	1.101	0.91 1.34	53.06	0.000
		Χρόνια Νεφρική Νόσος	7	1.162	0.83 1.64	10.41	0.109
		Νεφροπάθεια IgA	5	0.990	0.84 1.17	2.31	0.678
		Κυστεοουρητηρική Παλινδρόμηση	3	1.066	0.68 1.67	2.29	0.318
CC *vs* AA+AC		Νεφρική νόσος Τελικού Σταδίου	14	1.314	0.83 2.07	30.80	0.004
		Χρόνια Νεφρική Νόσος	6	1.059	0.50 2.25	3.16	0.675
		Νεφροπάθεια IgA	5	0.947	0.62 1.45	0.51	0.973
		Κυστεοουρητηρική Παλινδρόμηση	2	0.142	0.02 1.22	0.10	0.749
CC+AC *vs* AA		Νεφρική νόσος Τελικού Σταδίου	15	1.150	0.92 1.44	38.42	0.000
		Χρόνια Νεφρική Νόσος	7	1.159	0.82 1.63	11.03	0.087
		Νεφροπάθεια IgA	5	0.999	0.81 1.23	3.05	0.550
		Κυστεοουρητηρική Παλινδρόμηση	2	1.146	0.66 2.01	0.36	0.551
A *vs* C	A1332G/AGTR2	Κυστεοουρητηρική Παλινδρόμηση	4	1.018	0.668 1.551	5.44	0.143

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 2.6. Γραφική απεικόνιση του αποτελέσματος της μετα-ανάλυσης του πολυμορφισμού A1166C του γονιδίου AGTR1 (C *vs* A) και της συσχέτισής του με τη νεφρική νόσο τελικού σταδίου (επάνω) και τη χρόνια νεφρική νόσο(κάτω).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 2.12. Αποτελέσματα πολυπαραγοντικής μετα-ανάλυσης του πολυμορφισμού A1166C του γονιδίου AGTR1.

Ασθένεια	Αριθμός μελετών	Συνδυασμός Γονότυπων (Contrast)	OR	95% Διάστημα Εμπιστοσύνης	
Χρόνια Νεφρική Νόσος	6	AC vs AA	1.17	0.73	1.88
		CC vs AA	1.08	0.46	2.52
Νεφρική νόσος Τελικού Σταδίου	14	AC vs AA	1.14	0.94	1.37
		CC vs AA	1.29	0.78	2.15
Νεφροπάθεια IgA	5	AC vs AA	1.01	0.81	1.25
		CC vs AA	0.95	0.61	1.47
Κυστεοουρητηρική Παλινδρόμηση	2	AC vs AA	1.29	0.73	2.29
		CC vs AA	0.16	0.02	1.39

2.6.4 Συμπεράσματα

Τα αποτελέσματά μας υποδηλώνουν ότι ο πολυμορφισμός A1166C του γονιδίου AGTR1 δε συσχετίζεται με οποιαδήποτε από τις νεφρικές παθήσεις που εξετάστηκαν δηλαδή τη χρόνια νεφρική νόσο, τη νεφρική νόσο τελικού σταδίου, τη νεφροπάθεια IgA και την κυστεοουρητηρική παλινδρόμηση. Επιπλέον, ούτε ο πολυμορφισμός A1332G του γονιδίου AGTR2 συσχετίζεται με την κυστεοουρητηρική παλινδρόμηση. Τα αποτελέσματα αυτά είναι αντίθετα με τα στοιχεία άλλων μελετών που δείχνουν ότι πολυμορφισμοί άλλων γονιδίων (AGT και ACE) που εμπλέκονται στην ίδια βιοχημική οδό ρενίνης-αγγειοτενσίνης, σχετίζονται με τη νεφροπάθεια τελικού σταδίου (Zhou et al., 2013, 2014). Οι μη στατιστικά σημαντικές συσχετίσεις που βρέθηκαν μπορούν να αποδοθούν στο μικρό μέγεθος του δείγματος όλων των μετα-αναλύσεων που πραγματοποιήθηκαν. Σε μια προσπάθεια να εκτιμηθεί ο αριθμός των ανθρώπων που χρειάζονται επιπλέον για να ανιχνευθούν στατιστικά σημαντικές συσχετίσεις χρησιμοποιήθηκε η μέθοδος του Barrowman και βρέθηκε οτι χρειάζονται επιπλέον 4 έως 1500 άνθρωποι ανάλογα με την ασθένεια. Περισσότερες μελέτες τόσο γενετικής συσχέτισης όσο και ευρυγονιδιωματικές μελέτες είναι αναγκαίες προκειμένου να βρεθεί ο ρόλος των πολυμορφισμών των γονιδίων AGTR1 και AGTR2 στις νεφρικές παθήσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

2.7 Δημοσιεύσεις σε επιστημονικά περιοδικά

Τα αποτελέσματα που περιγράφηκαν στην ενότητα 2.4, έχουν δημοσιευτεί στο διεθνές επιστημονικό περιοδικό *International Journal of Molecular Sciences* (Braliou et al., 2015). Τα αποτελέσματα που περιγράφηκαν στην ενότητα 2.5 έχουν δημοσιευτεί στο διεθνές επιστημονικό περιοδικό *Psychiatric genetics* (Pantavou et al., 2016). Τα αποτελέσματα που περιγράφηκαν στην ενότητα 2.6 έχουν δημοσιευτεί στο διεθνές επιστημονικό περιοδικό *Computational and Structural Biotechnology Journal* (Braliou et al., 2014).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

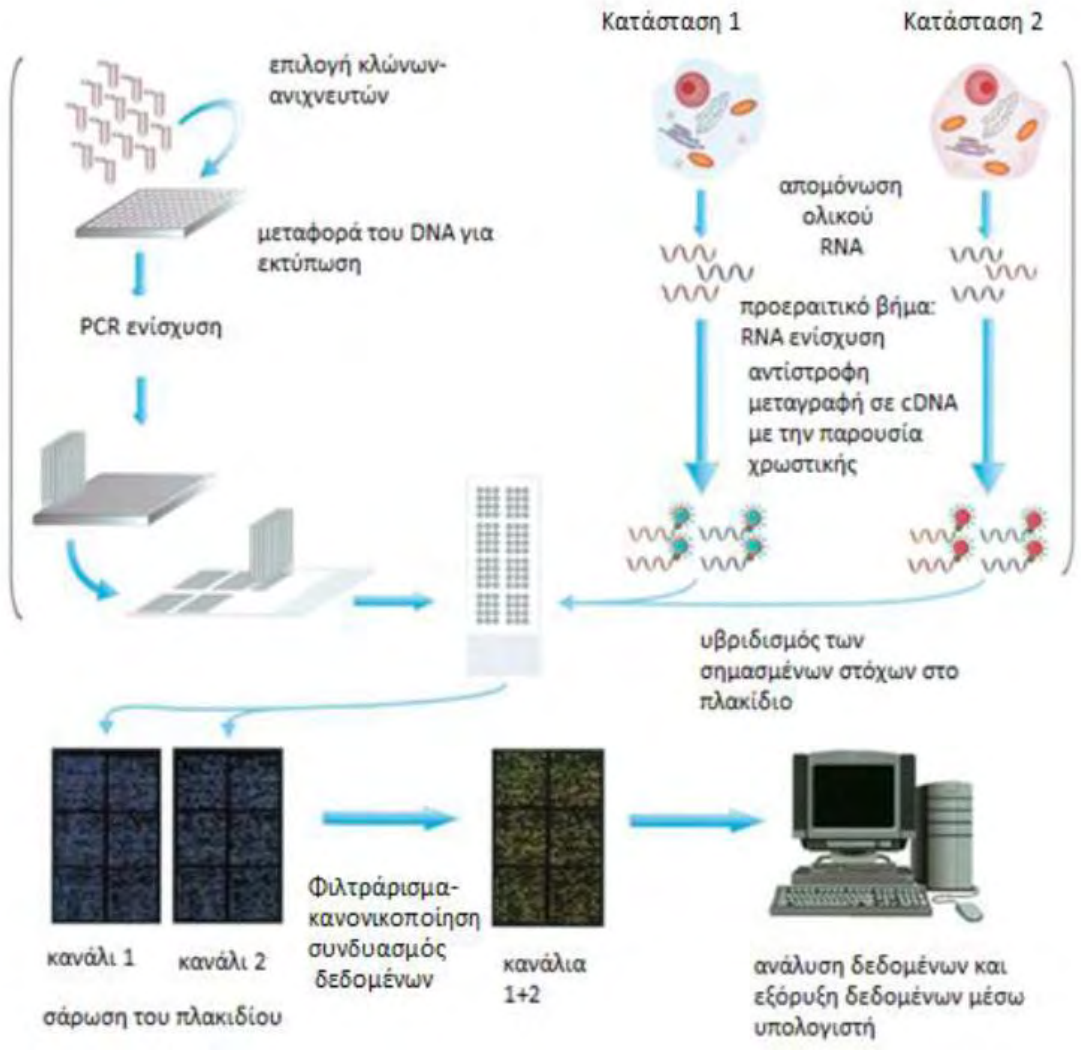
Κεφάλαιο 3

Μεθοδολογία ανάλυσης δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA

Η γονιδιακή έκφραση ασχολείται με την ποσότητα mRNA ή πρωτεϊνών που παράγονται από ένα κύτταρο σε μια δεδομένη στιγμή. Η ανάλυση της γονιδιακής έκφρασης βασίζεται στη σύγκριση δειγμάτων, για παράδειγμα ιστών –υγιών και ασθενών– για τη μελέτη συγκεκριμένων ασθενειών. Μέχρι το 1990, οι επιστήμονες μπορούσαν να μελετήσουν λίγα μόνο γονίδια κάθε φορά. Στις αρχές τις δεκαετίας του ’90 αναπτύχθηκε μία νέα τεχνολογία, η μικροσυστοιχία DNA -ή διαφορετικά, το DNA τσιπ-, η οποία επιτρέπει την ταυτόχρονη ανάλυση της έκφρασης χιλιάδων γονιδίων γρήγορα και αποτελεσματικά (Novoradovskaya et al., 2004). Μία από τις πιο διαδεδομένες βιολογικές εφαρμογές αυτής της τεχνολογίας είναι η σύγκριση των επιπέδων έκφρασης ενός συνόλου γονιδίων τα οποία διατηρούνται υπό συγκεκριμένες συνθήκες (κατάσταση Α) με τα επίπεδα έκφρασης ίδιου συνόλου γονιδίων τα οποία προέρχονται από ένα κύτταρο αναφοράς το οποίο διατηρείται υπό φυσιολογικές συνθήκες (κατάσταση Β), π.χ. η σύγκριση υγιών κυττάρων και κυττάρων τα οποία νοσούν έτσι ώστε να διερευνηθούν οι αιτίες οι οποίες προκαλούν τη νόσο. Άλλες εφαρμογές των μικροσυστοιχιών είναι στην εύρεση λειτουργιών νέων γονιδίων, αφού γονίδια με παρόμοιο τρόπο έκφρασης σε διαφορετικές πειραματικές συνθήκες μπορεί να εμπλέκονται στη διερεύνηση του τρόπου δράσης φαρμακευτικών σκευασμάτων και στη διευκόλυνση ταξινόμησης ασθενών (Brazma and Vilo, 2000).

Ένα τσιπ μικροσυστοιχιών αποτελείται από συγκεκριμένες αλληλουχίες οι οποίες είναι ειδικές για συγκεκριμένα γονίδια, τους ανιχνευτές (probes), οι οποίοι είναι ακινητοποιημένοι σε μία κουκκίδα (spot) μιας στερεάς επιφάνειας η οποία είναι συνήθως ένα πλακίδιο (chip) από γυαλί. Αυτές οι αλληλουχίες, στη συνέχεια, υποβάλλονται σε υβριδοποίηση με αντίγραφα νουκλεϊκών οξέων από βιολογικά δείγματα τα οποία είναι ιχνηθετημένα με φθορίζουσες ουσίες, τους στόχους (targets) (Lee et al., 2004). Χάρη στο φθορισμό, η ένταση της υβριδοποίησης, η οποία είναι ανάλογη του αριθμού των αντιγράφων κάθε είδους RNA το οποίο υπάρχει στο δείγμα, μπορεί να μετρηθεί με έναν σαρωτή με λέιζερ και να μετατραπεί σε μία ποσοτική τελική έξοδο. Με αυτόν τον τρόπο οι μικροσυστοιχίες επιτρέπουν την ταυτόχρονη μέτρηση των επιπέδων γονιδιακής έκφρασης χιλιάδων γονιδίων σε ένα και μόνο πείραμα υβριδοποίησης (Liu et al., 2007a).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 3.1. Ένα ολοκληρωμένο πείραμα μικροσυστοιχιών. Απομονώνεται βιολογικό υλικό (mRNA) από δύο τύπους κυττάρων, σημαίνεται με διαφορετικές χρωστικές και υβριδοποιείται στο πλακίδιο (chip). Τέλος, σαρώνεται και εξάγεται εικόνα προς επεξεργασία.

3.1 Διαδικασία πειράματος μικροσυστοιχιών

Η πειραματική διαδικασία αφορά τα στάδια τα οποία πρέπει να ακολουθηθούν κατά τη διεξαγωγή ενός πειράματος μικροσυστοιχιών. Ένα πείραμα με μικροσυστοιχίες είναι μια πολύπλοκη ακολουθία διεργασιών, οι οποίες πρέπει να εκτελεστούν με επιτυχία για να εξασφαλιστεί η εγκυρότητα των αποτελεσμάτων που θα προκύψουν.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Βήμα 1. Διατύπωση βιολογικού ερωτήματος

Το βιολογικό ερώτημα είναι ανάγκη να διατυπώνεται με ακρίβεια πριν προχωρήσει κάποιος στη διαδικασία ενός πειράματος με μικροσυστοιχίες. Ένα απλό ερώτημα θα αφορούσε τη διαφοροποίηση της γονιδιακής έκφρασης πριν και μετά τη χορήγηση ενός φαρμάκου. Σημαντικό ρόλο στις αποφάσεις για την εξέλιξη του πειράματος, παίζει το αν το βιολογικό υλικό είναι διαθέσιμο και αν η ποσότητά του είναι μικρή ή μεγάλη, καθώς αυτό θα επηρεάσει και την ένταση του φθορισμού και συνεπώς το αποτέλεσμα του πειράματος.

Βήμα 2. Επιλογή τύπου τσιπ μικροσυστοιχίας

Ανάλογα με το βιολογικό ερώτημα πρέπει να επιλεγεί και ο τύπος της μικροσυστοιχίας που θα χρησιμοποιηθεί στο πείραμα. Υπάρχουν δύο κύριοι τύποι μικροσυστοιχιών: α) οι μικροσυστοιχίες που κατασκευάστηκαν με *in situ* σύνθεση ανιχνευτών και β) οι μικροσυστοιχίες στις οποίες οι ανιχνευτές εκτυπώθηκαν στην επιφάνεια μετά την παρασκευή τους (Liu et al., 2007a). Η δεύτερη κατηγορία χωρίζεται, επίσης, σε δύο υποκατηγορίες σχετικές με το μήκος των ανιχνευτών. Έτσι, υπάρχουν μικροσυστοιχίες ολιγονουκλεοτιδίων και μικροσυστοιχίες cDNA.

Βήμα 3. Παρασκευή δείγματος

Σε ένα πείραμα μικροσυστοιχιών, το πρώτο στάδιο είναι η απομόνωση του βιολογικού υλικού το οποίο θα χρησιμοποιηθεί σαν στόχος. Στη συνέχεια, το mRNA το οποίο έχει απομονωθεί ιχνηθετείται. Η πιο απλή μέθοδος ιχνηθέτησης περιλαμβάνει τη χρήση νουκλεοτιδιών, τα οποία έχουν ήδη ιχνηθετηθεί με φθορίζουσα ουσία, στη διαδικασία της αντίστροφης μεταγραφής από mRNA σε cDNA. Στις τυπωμένες cDNA και ολιγονουκλεοτιδικές μικροσυστοιχίες χρησιμοποιούνται πάντα δύο δείγματα: ένα δείγμα αναφοράς και ένα δείγμα ελέγχου. Τα δείγματα mRNA ελέγχου και αναφοράς (για παράδειγμα, ο ιστός του ασθενή και ένας υγιής ιστός) μεταγράφονται σε cDNA με τη χρήση του ενζύμου αντίστροφη μεταγραφάση, ενώ ταυτοχρόνως γίνεται ιχνηθέτηση των μορίων cDNA, σε ξεχωριστούς σωλήνες, με τη χρήση δύο διαφορετικών ιχνηθετών (Freeman et al., 2000). Οι πιο δημοφιλείς επιλογές φθορίζουσων ουσιών είναι η Cy3 (πράσινο) και η Cy5 (κόκκινο). Το Cy5 χρησιμοποιείται συνήθως για την ιχνηθέτηση του δείγματος ελέγχου, ενώ το Cy3 για την ιχνηθέτηση του δείγματος αναφοράς. Τέλος, τα μόρια cDNA μεταφέρονται στις διατάξεις των μικροσυστοιχιών με σκοπό να υβριδοποιηθούν με τους ανιχνευτές. Οι σχετικές ποσότητες του κάθε μεταγράφου των γονιδίων στα δύο δείγματα εκτιμώνται μετρώντας την ένταση του σήματος φθορισμού από τα φάσματα εκπομπής, τα οποία προκύπτουν από τη διέγερση των ιχνηθετών στα αντίστοιχα μήκη κύματος αυτών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Βήμα 4. Υβριδοποίηση

Για να πραγματοποιηθεί η υβριδοποίηση του ιχνηθετημένου στόχου με τον αντίστοιχο ανιχνευτή, τοποθετούνται οι στόχοι, οι οποίοι βρίσκονται σε κατάλληλο ρυθμιστικό διάλυμα, πάνω στη διάταξη της μικροσυστοιχίας. Στη συνέχεια, οι στόχοι με τους ανιχνευτές επωάζονται για ένα συγκεκριμένο χρονικό διάστημα και σε μία συγκεκριμένη θερμοκρασία. Μετά την υβριδοποίηση, η οποία μπορεί να διαρκέσει από λίγες μέχρι αρκετές ώρες, το διάλυμα υβριδοποίησης αποβάλλεται και γίνονται πλύσεις των επιφανειών των μικροσυστοιχιών για την απομάκρυνση μη ειδικών υβριδοποιήσεων οι οποίες έχουν τυχόν πραγματοποιηθεί. Ύστερα, το πλακίδιο είναι έτοιμο για να σαρωθεί από ειδικό μηχάνημα.

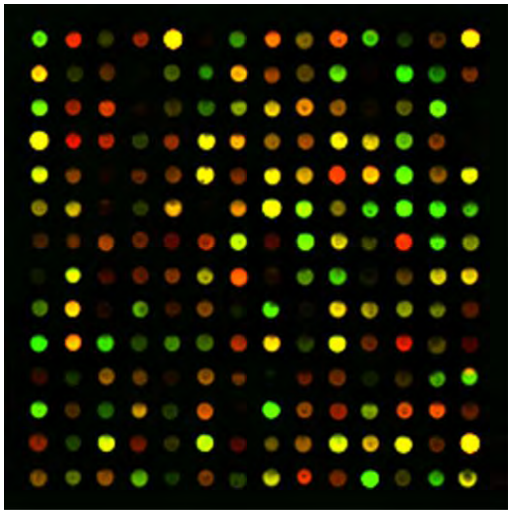
Βήμα 5. Σάρωση

Μετά από την υβριδοποίηση των δύο δειγμάτων (ελέγχου και αναφοράς) με τους ανιχνευτές, η πληροφορία των πειραμάτων των μικροσυστοιχιών μετατρέπεται σε εικόνες σαρώνοντας τα πλακίδια στα μήκη κύματος των δύο φθορίζουσων ουσιών και μετρώντας τον αντίστοιχο φθορισμό της κάθε ουσίας. Η αναλογία των εντάσεων φθορισμού για την κάθε κουκκίδα είναι ενδεικτική της σχετικής πληθώρας της αντίστοιχης ακολουθίας του DNA στα δύο δείγματα. Στην Εικόνα 3.2, μια κουκκίδα εμφανίζεται με χρώμα αντίστοιχο με την ποσότητα του δείγματος ελέγχου και του δείγματος αναφοράς. Έτσι, προκύπτουν οι εξής ερμηνείες για τα χρώματα των κουκκίδων:

- Με κόκκινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος ελέγχου είναι μεγαλύτερη.
- Με πράσινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος αναφοράς είναι μεγαλύτερη.
- 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). Στη συνέχεια, συλλέγονται οι πληροφορίες σχετικά με την έκφραση των γονιδίων της μικροσυστοιχίας με τη βοήθεια ενός λογισμικού επεξεργασίας και ανάλυσης εικόνων μικροσυστοιχιών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 3.2. Η εικόνα της μικροσυστοιχίας

Βήμα 6. Κανονικοποίηση

Έχοντας ως σκοπό την ακριβή μέτρηση των επιπέδων έκφρασης των γονιδίων, είναι σημαντικό να λάβουμε υπόψη τα τυχαία και τα συστηματικά σφάλματα τα οποία συμβαίνουν σε ένα πείραμα μικροσυστοιχιών. Για παράδειγμα, μια γνωστή και συνήθης πηγή συστηματικού σφάλματος προκύπτει από τη χρήση διαφορετικών φθορίζουσων ουσιών. Αυτό το φαινόμενο γίνεται φανερό αν πραγματοποιήσουμε ένα πείραμα, όπου δύο πανομοιότυπα mRNA δείγματα ιχνηθέτουνται με διαφορετική χρωστική (με Cy3 το ένα και με Cy5 το άλλο) και στη συνέχεια υβριδοποιούνται πάνω στο ίδιο πειραματικό πλακίδιο. Είναι ιδιαίτερα σπάνιο να έχουμε ίσες εντάσεις για τις δύο χρωστικές ουσίες στις κουκκίδες του πλακιδίου, και μάλιστα, συνήθως έχουμε υψηλότερα επίπεδα έκφρασης για τη Cy3 από αυτά της Cy5. Αν και τέτοιου είδους συστηματικά σφάλματα έχουν σχετικά μικρή τιμή, μπορούν να δημιουργήσουν μεγάλο πρόβλημα όσον αφορά την αξιοπιστία των αποτελεσμάτων, όπως κατά την αναζήτηση μικρών βιολογικών διαφορών. Τα σφάλματα μπορούν να προέρχονται από διάφορους παράγοντες, συμπεριλαμβανομένων των φυσικών ιδιοτήτων των φθορίζουσών ουσιών, της απόδοσης ενσωμάτωσης των χρωστικών, της μεταβλητότητας στην πειραματική διαδικασία κατά την υβριδοποίηση και των μετέπειτα διαδικασιών επεξεργασίας, ακόμη και των ρυθμίσεων του σαρωτή. Επιπροσθέτως, τα σχετικά επίπεδα έκφρασης των γονιδίων από επαναλήψεις του ίδιου πειράματος μπορεί να έχουν διαφορετική διακύμανση λόγω διαφορετικών πειραματικών συνθηκών. Πολλοί από τους παράγοντες οι οποίοι προκαλούν αυτά τα συστηματικά σφάλματα, είναι είτε εσωτερικοί είτε εξωτερικοί της υβριδοποίησης των δειγμάτων στο τσιπ και πρέπει να ελαχιστοποιηθούν όσο το δυνατό με τη βοήθεια της κανονικοποίησης προκειμένου να πάρουμε αξιόπιστα αποτελέσματα. Κατά συνέπεια,

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

σκοπός της κανονικοποίησης (normalization) είναι η ελαχιστοποίηση των συστηματικών σφαλμάτων τα οποία εντοπίζονται στα εκτιμώμενα επίπεδα έκφρασης των γονιδίων, έτσι ώστε οι βιολογικές διαφορές να γίνονται ευκολότερα αντιληπτές και να επιτρέπεται η σύγκριση των επιπέδων έκφρασης μεταξύ δεδομένων τα οποία προέρχονται από διαφορετικά τσιπ μικροσυστοιχιών (Quackenbush, 2001).

Βήμα7. Στατιστική Ανάλυση

Τα δεδομένα από το τσιπ των μικροσυστοιχιών αποτελούν έναν πίνακα γονιδιακής έκφρασης. Σε αυτόν τον πίνακα, οι γραμμές αναπαριστούν τα γονίδια και οι στήλες τα δείγματα. Οι τιμές σε κάθε θέση αυτού του πίνακα αντιστοιχούν στις τιμές έκφρασης (expression values) του συγκεκριμένου γονιδίου κάτω από τις συγκεκριμένες πειραματικές συνθήκες. Συνήθως αυτές οι τιμές είναι ο λογάριθμος με βάση το 2 του λόγου της έντασης για το κόκκινο κανάλι προς την ένταση για το πράσινο κανάλι. Η πολυπλοκότητα των αποτελεσμάτων τα οποία παράγονται από τα πειράματα με μικροσυστοιχίες από ένα σχετικά μικρό αριθμό δειγμάτων επιβάλει τη χρήση στατιστικών εργαλείων για την ανάλυσή τους. Στόχος της ανάλυσης αυτής είναι η εύρεση πιθανών σχέσεων που εμφανίζουν τα δεδομένα. Για παράδειγμα, ενδιαφέρον έχει η εύρεση γονιδίων τα οποία εκφράζονται διαφορικά σε μια ασθένεια με στόχο να χρησιμοποιηθούν ως προγνωστικός δείκτης για την ασθένεια αυτή. Για την ανάλυση δεδομένων μικροσυστοιχιών μπορεί να χρησιμοποιηθεί μια πληθώρα στατιστικών μεθόδων, οι οποίες θα περιγραφούν στην επόμενη ενότητα, που θα περιλαμβάνει και παρουσίαση μιας συγκριτικής μελέτης όλων των προτεινόμενων μεθόδων.

3.2 Μέθοδοι Ανάλυσης Μικροσυστοιχιών

Οι περισσότερες μελέτες μικροσυστοιχιών αξιολογούν τη διαφορική έκφραση μόνο από την άποψη της αλλαγής του διπλώματος (Fold Change/ FC), με τιμές FC ± 2 να θεωρούνται και τα γονίδια με αυτές τις τιμές να εκφράζονται διαφορικά. Ωστόσο, οι τιμές τις αλλαγής του διπλώματος (FC) δεν λαμβάνουν υπόψη τη μεταβλητότητα των δεδομένων μεταξύ ενός πειράματος και επιπλέον δε μπορεί να εξασφαλιστεί η επαναληψιμότητα. Επιπλέον, η κατάταξη με βάση την αλλαγή διπλώματος δεν είναι επαρκής, δεδομένου ότι ένα γονίδιο με μεγαλύτερη διακύμανση στις τιμές έκφρασης έχει μεγαλύτερη πιθανότητα να είναι στατιστικά σημαντικό. Οι μέθοδοι βασισμένες στην αλλαγή διπλώματος χρησιμοποιούνται συνηθώς σε περιπτώσεις που είναι εύκολο να δημιουργηθούν σύνολα διαφορικά εκφρασμένων γονιδίων τα οποία είναι εύκολο να αναπαραχθούν. Ωστόσο, η επαναληψιμότητα δεν συνεπάγεται και την ακρίβεια και το ζήτημα του αν θα χρησιμοποιηθεί η αλλαγή διπλώματος για να

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

βρεθούν τα διαφορικά εκφρασμένα γονίδια είναι κυρίως βιολογικό παρά στατιστικό πρόβλημα (Witten and Tibshirani, 2007).

3.2.1 Έλεγχος *t* (*t*-test)

Το *t*-test αξιολογεί αν οι μέσες τιμές μεταξύ δύο ομάδων έχουν στατιστικά σημαντικές διαφορές. Το *t*-test ενός δείγματος (one sample *t*-test) είναι μια στατιστική μέθοδος που χρησιμοποιείται για τον προσδιορισμό της διαφοράς μέσων τιμών μεταξύ των δειγμάτων με γνωστή ή υποθετικά γνωστή τη μέση τιμή του πληθυσμού. Η μηδενική υπόθεση υποθέτει ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των δειγμάτων και της μέσης τιμής του πληθυσμού. Σε πραγματικές εφαρμογές το *t*-test ενός δείγματος χρησιμοποιείται για τον έλεγχο στατιστικών διαφορών σε ζεύγη παρατηρήσεων, είτε μετρώντας το ίδιο δείγμα (για παράδειγμα, ένα άτομο πριν και μετά την θεραπεία) ή, γενικότερα, όταν τα δείγματα με κάποιο τρόπο ομαδοποιούνται (ζευγαρωτές παρατηρήσεις):

$$\bar{X}_1 - \bar{X}_2 = \bar{X}_D \qquad (1)$$

Για τον υπολογισμό της τιμής του t-test ενός δείγματος, χρησιμοποιούμε τον τύπο:

$$t = \frac{\bar{X}_D}{S_D / \sqrt{n}} \qquad (2)$$

Σε αυτή την ανάλυση, οι δείκτες 1 και 2 παριστάνουν τις δύο καταστάσεις, η διαφορά των μέσων τιμών υποτίθεται ότι είναι μηδέν, σύμφωνα με τη μηδενική υπόθεση. Ο εκτιμητής *t* συγκρίνεται με την κατανομή *t* με *n*-1 βαθμούς ελευθερίας.

Το *t*-test δύο δειγμάτων (two sample *t*-test) χρησιμοποιείται για να συγκρίνει τις μέσες τιμές δυο ανεξάρτητων πληθυσμών. Η μηδενική υπόθεση υποθέτει ότι οι μέσες τιμές των δύο πληθυσμών είναι ίσες. Ο εκτιμητής *t* υποθέτοντας ίσες διακυμάνσεις μπορεί να υπολογιστεί ως εξής:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3)$$

όπου η διακύμανση S_p δίνεται από την εξίσωση:

$$S_p = \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}} \qquad (4)$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Ο εκτιμητής *t* συγκρίνεται με την κατανομή *t* με n_1+n_2-2 βαθμούς ελευθερίας. Στην περίπτωση που οι διακυμάνσεις είναι άνισες, προκειμένου να ελεγχθεί η ισότητα του πληθυσμού ο εκτιμητής *t* υπολογίζεται ως εξής:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \quad (5)$$

με $S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ (6)

Ωστόσο, στην περίπτωση αυτή η ασυμπτωτική κατανομή είναι δύσκολο να βρεθεί και μία προσέγγιση είναι να χρησιμοποιηθεί μια *t*-κατανομή με βαθμούς ελευθερίας οι οποίοι δίνονται από τον τύπο:

$$d.f. = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1 - 1) + \left(S_2^2/n_2\right)^2/(n_2 - 1)} \quad (7)$$

Ένα σημαντικό μειονέκτημα του ελέγχου *t* για την ανάλυση δεδομένων των μικροσυστοιχιών είναι ότι τα περισσότερα πειράματα μικροσυστοιχιών περιέχουν λίγα δείγματα σε κάθε ομάδα (n_1 και n_2) οπότε δεν ισχύει η υπόθεση της κανονικότητας. Έτσι, αρκετές εναλλακτικές μέθοδοι έχουν προταθεί στη βιβλιογραφία για τον έλεγχο *t*.

3.2.2 Μέθοδοι Αναδειγματοληψίας (Resampling methods)

Η αναδειγματοληψία Bootstrap (Efron, 1982; Efron and Tibshirani, 1993) είναι μια στατιστική μέθοδος για την εκτίμηση της κατανομής του εκτιμητή με δειγματοληψία με αντικατάσταση από το αρχικό δείγμα. Η Bootstrap αποτελεί μια ιδανική εναλλακτική μέθοδο όταν δεν είναι διαθέσιμη η δειγματική κατανομή ή όταν είναι διαθέσιμοι οι μαθηματικοί τύποι αλλά δεν ικανοποιούνται οι κατάλληλες υποθέσεις (π.χ. μικρό μέγεθος δείγματος ή μη-κανονική κατανομή). Η ακρίβεια της Bootstrap εξαρτάται από τον αριθμό των παρατηρήσεων στο αρχικό δείγμα και τον αριθμό των επαναλήψεων. Μια μικρή δειγματοληψία είναι επαρκής για να υπολογίσει το τυπικό σφάλμα αλλά χρειάζεται μεγαλύτερο δείγμα για την κατασκευή ενός 95% διαστήματος εμπιστοσύνης. Υπάρχουν διάφορες μέθοδοι για την κατασκευή ενός διαστήματος εμπιστοσύνης από Bootstrap, όπως είναι η κανονική μέθοδος προσέγγισης (normal approximation method), η μέθοδος διόρθωσης του σφάλματος (the bias corrected method), η μέθοδος ποσοστημορίων (percentile method) και η μέθοδος ποσοστημορίου *t* (*t*-percentile method) (Efron, 1987). Γενικά, επαναλήψεις της τάξης των 1.000 παράγουν πολύ ακριβείς

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

εκτιμήσεις, αν και μπορεί να χρειαστούν περισσότερες για την ακριβή εκτίμηση του p-value. Μόνο 50-200 επαναλήψεις απαιτούνται για την εκτίμηση των τυπικών σφαλμάτων, αν και αυτό μπορεί να έχει επιπτώσεις στις μεθόδους μετα-ανάλυσης (βλέπε Κεφάλαιο 4). Διάφορες μέθοδοι έχουν προταθεί για την εκτίμηση του επαρκούς αριθμού επαναλήψεων (Andrews and Buchinsky, 2000; Davidson and MacKinnon, 2000). Η Bootstrap έχει εφαρμοστεί σε πειράματα μικροσυστοιχιών και εμπειρικά στοιχεία δείχνουν ότι παράγει ακριβείς εκτιμήσεις, τουλάχιστον για μέτρια μεγέθη δειγμάτων (Meuwissen and Goddard, 2004). Για πραγματικά μικρά μεγέθη δείγματος (δηλαδή <10), διάφορες τροποποιήσεις στην τυποποιημένη μέθοδο Bootstrap έχουν προταθεί (Jiang and Simon, 2007; Neuhauser and Jockel, 2006).

Μια διαφορετική μέθοδος αναδειγματοληψίας είναι ο έλεγχος μετάθεσης (permutation test). Αυτό είναι ένα είδος ελέγχου της στατιστικής σημαντικότητας στην οποία η κατανομή του στατιστικού κάτω από τη μηδενική υπόθεση προκύπτει από τον υπολογισμό όλων των πιθανών τιμών του στατιστικού μετά από ανακατατάξεις των ετικετών των παρατηρήσεων. Αν κάτω από τη μηδενική υπόθεση οι ετικέτες είναι ανταλλάξιμες, οι προκύπτοντες έλεγχοι παράγουν ακριβή επίπεδα σημαντικότητας. Στη συνέχεια, υπολογίζονται τα διαστήματα εμπιστοσύνης από τους ελέγχους. Η θεωρία του ελέγχου μετάθεσης έχει αναπτυχθεί από τον Fisher και Pitman στη δεκαετία του 1930 και μια πρόσφατη ανασκόπηση έγινε από τον Kaiser (Kaiser, 2007). Για μικρά δείγματα, όλοι οι πιθανοί συνδυασμοί μπορούν να μετρηθούν. Ωστόσο, για μέγεθος δείγματος μεγαλύτερο από 15, ένα τυχαίο δείγμα μετάθεσης χρησιμοποιείται αντ 'αυτού, και γι’αυτό το λόγο προέκυψε το όνομα μετάθεση Μόντε Κάρλο. Μια σημαντική υπόθεση στην οποία βασίζεται ο έλεγχος μετάθεσης είναι ότι κάτω από την μηδενική υπόθεση οι παρατηρήσεις είναι ανταλλάξιμες. Συνέπεια αυτού είναι ότι οι έλεγχοι της διαφοράς των μέσων τιμών (π.χ. *t*-test) απαιτούν ίση διακύμανση. Από την άποψη αυτή, ο έλεγχος μετάθεσης *t* έχει την ίδια αδυναμία όπως ο κλασικός έλεγχος *t*. Γενικά, ο έλεγχος μετάθεσης υπολογίζει μια τιμή *p* (*p*-value) με απαρίθμηση των περιπτώσεων στις οποίες το στατιστικό είναι μεγαλύτερο από το παρατηρηθέν. Έτσι, απαιτείται ένας μεγάλος αριθμός επαναλήψεων της τάξης των 1000 ή και περισσότερο. Οι έλεγχοι μετάθεσης έχουν χρησιμοποιηθεί για την ανάλυση των δεδομένων των μικροσυστοιχιών (Tsai et al., 2003). Στην περίπτωση που το μέγεθος του δείγματος είναι πολύ μικρό, ο αριθμός των διακριτών μεταθέσεων μπορεί να περιοριστεί σημαντικά και έχει προταθεί ο συνδυασμός στατιστικών μεθόδων μετάθεσης για κάθε γονίδιο. Ωστόσο, δεδομένου ότι η κατανομή στη μηδενική υπόθεση των στατιστικών υπό μετάθεση δεν είναι η ίδια για όλα τα γονίδια, αυτό μπορεί να έχει αρνητικό αντίκτυπο στην εκτίμηση του p-value (Yang and Churchill, 2007).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Η μέθοδος Bootstrap και η μέθοδος μετάθεσης είναι άμεσα διαθέσιμες στα κυριότερα στατιστικά πακέτα όπως το Stata (StataCorp, 2013) και η R (R Core Team, 2016). Υπάρχουν διάφορες υλοποιήσεις της μεθόδου Bootstrap στο Stata (εντολή *Bootstrap)* και στην R (εντολή *boot*). Ο έλεγχος με μετάθεση μπορεί να γίνει επίσης με τη χρήση των εντολών *permute* και *permtest* (για ζεύγη παρατηρήσεων) στο Stata, καθώς και με την εντολή *perm* στο R. Στην παρούσα διατριβή υλοποιήθηκαν δυο προγράμματα που εφαρμόζουν τον έλεγχο Bootstrap και τον έλεγχο μετάθεσης για δεδομένα από μικροσυστοιχίες στο στατιστικό πακέτο Stata. Όλα τα απαραίτητα στοιχεία που χρειάζονται για την εκτέλεση των εντολών αυτών βρίσκονται διαθέσιμα στο ευρύ κοινό μέσω της ιστοσελίδας www.compgen.org/tools/microarrays.

3.2.3 Μπεϋζιανές Μέθοδοι (Bayesian methods)

Οι Μπεϋζιανές μέθοδοι παρέχουν ένα ελκυστικό πλαίσιο σχετικά με το χειρισμό των περισσότερων προβλημάτων που ανακύπτουν κατά την ανάλυση των δεδομένων των μικροσυστοιχιών. Αρκετές Μπεϋζιανές μέθοδοι έχουν αναπτυχθεί για να αντικαταστήσουν τον έλεγχο *t*, ο οποίος είναι μια από τις απλούστερες και ευρέως χρησιμοποιούμενες στατιστικές μεθόδους στην ανάλυση των δεδομένων έκφρασης από μικροσυστοιχίες. Οι διάφορες Μπεϋζιανές μέθοδοι έχουν κάποια κοινά χαρακτηριστικά μεταξύ τους, αλλά έχουν επίσης σημαντικές διαφορές ανάλογα με τα κριτήρια που χρησιμοποιούν, ιδιαίτερα στη γνώση της εκ των προτέρων κατανομής (prior distribution) των υπερπαραμέτρων. Επιπλέον, ορισμένες από τις μεθόδους αυτές είναι προσανατολισμένες ως προς τον έλεγχο υποθέσεων στηριζόμενες στον παράγοντα Bayes για να συγκρίνουν τη μηδενική έναντι της εναλλακτικής υπόθεσης (Gönen et al., 2005; Gottardo et al., 2003; Rouder et al., 2009; Wang and Liu, 2015). Άλλες μέθοδοι προσανατολίζονται προς την εκτίμηση των παραμέτρων που μας ενδιαφέρουν και υπολογίζουν διαστήματα αξιοπιστίας, όπως για παράδειγμα για τη διαφορά των μέσων τιμών (Kruschke, 2013; Wetzels et al., 2009). Ένα από τα πλεονεκτήματα του ελέγχου *t* είναι η απλότητά του, η οποία επιτρέπει σε πολλές περιπτώσεις τον υπολογισμό μιας έκφρασης κλειστής μορφής, ειδικά για τον παράγοντα Bayes (Gönen et al., 2005; Gottardo et al., 2003; Rouder et al., 2009; Wang and Liu, 2015). Ένα άλλο σημαντικό πλεονέκτημα των Μπεϋζιανών μεθόδων είναι ότι εντός του Μπεϋζιανού πλαισίου, δεν μπορεί κανείς να ενσωματώσει μόνο την αβεβαιότητα σχετικά με τις παραμέτρους και το μικρό μέγεθος του δείγματος, αλλά και των πολλαπλών ελέγχων, το οποίο είναι πολύ σημαντικό στην ανάλυση μικροσυστοιχιών (Fox and Dimmic, 2006; Gonen, 2010; Gottardo et al., 2003).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Υπάρχουν πολλές εφαρμογές λογισμικού που διατίθενται για τις προαναφερθείσες Μπεϋζιανές μεθόδους. Για παράδειγμα, η μέθοδος Bayes Factor των Rouder και συνεργάτων (Rouder et al., 2009), η οποία είναι γνωστή και ως έλεγχος *t* τύπου Jeffreys–Zellner–Siow (JZS), και είναι διαθέσιμη στην ιστοσελίδα: http://pcl.missouri.edu/ bayesfactor, καθώς και σε ένα πακέτο της R (https://cran.r-project.org/web/packages/BayesFactor/index.html). Ο έλεγχος *t* τύπου Savage-Dickey (SD) που προτάθηκε από τον Wetzels και τους συνεργάτες του (Wetzels et al., 2009) είναι εμπνευσμένος από τον έλεγχο *t*-JZS και διατηρεί τις βασικές του ιδιότητες. Χρησιμοποιείται όμως σε ένα πιο ευρύ φάσμα στατιστικών προβλημάτων διότι επιτρέπει στους ερευνητές να ελέγξουν αν υπάρχουν περιορισμοί και έχει εφαρμογή σε καταστάσεις δύο δειγμάτων με άνιση διακύμανση. Ο έλεγχος *t*-SD είναι διαθέσιμος σε ένα πακέτο της R που χρησιμοποιεί το πρόγραμμα WinBUGS (http://www.ruudwetzels.com/sdtest). Τέλος, η μέθοδος BEST (Μπεϋζιανή εκτίμηση που αντικαθιστά τον έλεγχο t/Bayesian Estimation Supersedes the *t*-test) παρέχει μια εναλλακτική λύση για Μπεϋζιανο έλεγχο *t* παρέχοντας πολύ πιο πλούσια πληροφορία από ένα απλό p-value, όπως πλήρη διαστήματα αξιοπιστίας για το μέγεθος επίδρασης, τη διαφορά της μέσης τιμής μεταξύ των ομάδων, τη διαφορά των τυπικών αποκλίσεων και την κανονικότητα των δεδομένων μεταξύ των ομάδων (Kruschke, 2013). Η μέθοδος BEST είναι υλοποιημένη στην R (http://www.indiana.edu/~kruschke/BEST/) και είναι διαθέσιμη διαδικτυακά στην ιστοσελίδα (http://sumsar.net/best_online/). Επιπλέον, η μέθοδος BEST υλοποιείται και στο πακέτο Bayesian First Aid package (https://github.com/rasmusab/bayesian_first_aid) που έχει ως στόχο να παρέχει φιλικές προς το χρήστη παραλλαγές των Μπεϋζιανών μεθόδων με τις πιο ευρέως χρησιμοποιούμενες εντολές για εκτίμηση των Μπεϋζιανών παραμέτρων.

3.2.4 Έλεγχος *t* με ποινή (Penalized t-test)

Όπως έχουμε ήδη αναφέρει, ο κλασικός έλεγχος *t* δεν συνιστάται για πειράματα μικροσυστοιχιών επειδή μια μεγάλη τιμή του εκτιμητή *t* μπορεί να προκύψει από μια εξωπραγματικά μικρή τιμή της διακύμανσης. Γονίδια με μικρές διακυμάνσεις, πιθανώς λόγο του πολύ μικρού μεγέθους δείγματος, έχουν μεγάλη πιθανότητα να δώσουν ένα μεγάλο εκτιμητή *t* ακόμη και αν δεν εκφράζονται διαφορικά. Εναλλακτικές μέθοδοι έχουν προταθεί, προκειμένου να αντιμετωπιστούν τα προβλήματα αυτά. Οι περισσότερες από αυτές τις μεθόδους χαρακτηρίζονται από μια εμπειρική Μπεϋζιανή αιτιολόγηση, και ως εκ τούτου μοιράζονται πολλά κοινά χαρακτηριστικά με τις Μπεϋζιανές μεθόδους. Άλλες μέθοδοι αποτελούνται κυρίως από ad-hoc τροποποιήσεις. Σε κάθε περίπτωση, όλες αυτές οι μέθοδοι

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

εφαρμόζουν διάφορα είδη τροποποίησης του παρονομαστή στον τύπο του ελέγχου *t* αυξάνοντας με αυτό τον τρόπο τη διακύμανση (Kooperberg et al., 2005). Έτσι, όλοι έχουν συνήθως την ίδια ερμηνεία με τον κλασικό έλεγχο *t*. Οι Baldi και Long ήταν οι πρώτοι οι οποίοι είχαν προτείνει Μπεϋζιανές μεθόδους στον έλεγχο *t* στο πλαίσιο των πειραμάτων μικροσυστοιχιών (Baldi and Long, 2001; Kayala and Baldi, 2012) αλλά προτίμησαν να αναπτύξουν μια εμπειρική Μπεϋζιανή μέθοδο ελέγχου *t* με διακύμανση ίση με:

$$S^2_{Cyber-T} = \frac{\nu_0\sigma_0^2 + (n-1)S^2}{\nu_0 + n - 2} \qquad (8)$$

Η μέθοδος αυτή είναι διαθέσιμη στην ιστοσελίδα Cyber-T (http://cybert.ics.uci.edu/) και στην R (http://cybert.ics.uci.edu/). Η παράμετρος ν_0 αντιπροσωπεύει το βαθμό εμπιστοσύνης στη διακύμανση σ_0^2 σε σχέση με την εμπειρική διακύμανση. Στην Cyber-T, η τιμή των ν_0 είναι καθορισμένη από το χρήστη. Όσο μικρότερη είναι η τιμή *n*, τόσο μεγαλύτερη είναι η τιμή ν_0. Ένας απλός κανόνας είναι να υποθέσουμε ότι K> 2 για να εκτιμηθεί σωστά η τυπική απόκλιση και να θέσουμε $n + \nu_0 = K$. Αυτό επιτρέπει μια πιο ευέλικτη αντιμετώπιση των καταστάσεων στις οποίες ο αριθμός *n* των διαθέσιμων δεδομένων ποικίλλει από γονίδιο σε γονίδιο. Η προεπιλεγμένη τιμή του K είναι 10. Συγκεκριμένα, με τη χρήση αυτής της προσέγγισης, η εμπειρική διακύμανση εξαρτάται από τις ν_0 "ψευδο-παρατηρήσεις" με διακύμανση σ_0^2. Για σ_0, κάποιος θα μπορούσε να χρησιμοποιήσει την τυπική απόκλιση ολόκληρου του συνόλου δεδομένων ή από συγκεκριμένες κατηγορίες γονιδίων. Η μέθοδος Cyber-T χρησιμοποιεί μια ευέλικτη προσέγγιση σύμφωνα με την οποία η τυπική απόκλιση εκτιμάται, με τη συγκέντρωση όλων των γειτονικών γονιδίων που περιέχονται σε ένα παράθυρο του μεγέθους *w* (η προεπιλεγμένη τιμή του *w* είναι 101, που αντιστοιχεί σε 50 γονίδια σε εκατέρωθεν του γονιδίου υπό ανάλυση).

Μια άλλη εμπειρική Μπεϋζιανή μέθοδος είναι η μέθοδος των Lönnstedt και Speed (Lönnstedt and Speed, 2002), η οποία χρησιμοποιεί την τροποποιημένη διακύμανση:

$$S^2_{LS} = a + S^2 \qquad (9)$$

στην οποία η ποινή α υπολογίζεται από τη μέση τιμή και την τυπική απόκλιση του δείγματος με διακυμάνσεις *S*. Αργότερα, ο Smyth (Smyth, 2005) πρότεινε μια παραλλαγή του τύπου αυτού, η οποία υλοποιείται στο γνωστό πακέτο ανάλυση δεδομένων μικροσυστοιχιών *limma*:

$$S^2_{limma} = \frac{\nu_0\sigma_0^2 + nS^2}{\nu_0 + n} \qquad (10)$$

Τα d_0 και s_0 υπολογίζονται από τα δεδομένα με τη μέθοδο των ροπών (method of moments) χρησιμοποιώντας μια εμπειρική Μπεϋζιανή προσέγγιση. Η μέθοδος limma είναι μία από τις πιο ευρέως

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

χρησιμοποιούμενες μεθόδους στην εύρεση διαφορικά εκφρασμένων γονιδίων και είναι διαθέσιμη ως πακέτο Bioconductor στην R (http://bioinf.wehi.edu.au/limma). Οι μέθοδοι των Tusher και συνεργατών (Tusher et al., 2001) και Efron και συνεργατών (Efron et al., 2001) χρησιμοποιούνται επίσης ως έλεγχοι *t* με ποινή:

$$S_{SAM} = a + S \qquad (11)$$

Η μέθοδος αυτή διαφέρει ελαφρώς από τις προηγούμενες στο γεγονός ότι η ποινή *α* εφαρμόζεται στην τυπική απόκλιση *S* και όχι στη διακύμανση S^2 του δείγματος. Ο Tusher οι και συνεργάτες του (Tusher et al., 2001) στη μέθοδο «Ανάλυση Σημαντικότητας των Μικροσυστοιχιών (Significance Analysis of Microarrays- SAM)», επέλεξαν το *α* έτσι ώστε να ελαχιστοποιηθεί ο συντελεστής διακύμανσης των απολύτων τιμών *t* ενώ ο Efron και οι συνεργάτες του (Efron et al., 2001) χρησιμοποίησαν το *α* ως το 90^o εκατοστημόριο των τιμών της τυπικής απόκλισης *S*. Αυτές οι επιλογές βασίζονται σε εμπειρικές και όχι σε θεωρητικές εκτιμήσεις. Η μέθοδος SAM είναι μια από τις παλαιότερες και ευρέως χρησιμοποιούμενες μεθόδους και είναι διαθέσιμη ως πρόσθετη εντολή στο Excel στη διεύθυνση http://statweb.stanford.edu/~tibs/SAM/, ενώ επιπλέον είναι υλοποιημένη σε διάφορα πακέτα της R (*samr, ema*).

Η μέθοδος Ανάλυσης Κατάταξης των Δεδομένων των Μικροσυστοιχιών (Ranking Analysis of Microarray data-RAM) χρησιμοποιεί ένα άλλο είδος κανονικοποίησης, η οποία βασίζεται στην παρατήρηση ότι, ακόμη και όταν το μέγεθος του δείγματος είναι μικρό, χρησιμοποιώντας τον τροποποιημένο έλεγχο *t*, το αποτέλεσμα μπορεί είναι ακόμα αρκετά ισχυρό (Tan et al., 2006). Ειδικότερα, το μικρό μέγεθος του δείγματος οδηγεί συχνά σε αδικαιολόγητα μεγάλη τιμή του *α* η οποία κυριαρχεί στον εκτιμητή *t* και κατά συνέπεια μειώνεται η ισχύς της ανάλυσης. Έτσι, οι συγγραφείς πρότειναν:

$$S_{RAM} = \begin{cases} 1 + S \ \ if \ \ \bar{X}_D > S < 1 \\ S \ \ otherwise \end{cases} \qquad (12)$$

Η μέθοδος RAM βασίζεται στις συγκρίσεις μεταξύ ενός συνόλου καταταγμένων εκτιμητών *t* και ένα σύνολο καταταγμένων Z-τιμών (ένα σύνολο καταταγμένων εκτιμήσεων για τη μηδενική υπόθεση) τα οποία προέκυψαν από μια προσέγγιση τυχαίου διαχωρισμού (“randomly splitting”), αντί της προσέγγισης με μετάθεση που χρησιμοποιείται από τη μέθοδο SAM. Τα αποτελέσματα που προκύπτουν από προσομοίωση και από πραγματικά δεδομένα μικροσυστοιχιών έδειξαν ότι η μέθοδος RAM είναι πιο αποτελεσματική στην εύρεση διαφορικά εκφρασμένων γονιδίων σε περιπτώσεις με μικρό μέγεθος

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

δείγματος, μεγάλο παράγοντα παραποίηση (fudge factor) ή ύπαρξης θορύβου, σε σύγκριση με τη μέθοδο SAM.

Οι παραλλαγές του ελέγχου *t* έχουν αρκετά πλεονεκτήματα. Πιο συγκεκριμένα, υπολογίζονται εύκολα, έχουν μια φυσική ερμηνεία, και είναι λιγότερο υπολογιστικά χρονοβόρες σε σύγκριση με τις Μπεϋζιανές μεθόδους και τις μεθόδους αναδειγματοληψίας. Επιπλέον, οι μελέτες προσομοίωσης (Witten and Tibshirani, 2007) έχουν δείξει ότι οι παραλλαγές του ελέγχου *t* είναι καλύτερες από τον κλασικό έλεγχο *t* για την ανίχνευση των διαφορικά εκφρασμένων γονιδίων, ακόμη και όταν το μέγεθος του δείγματος είναι πολύ μικρό (n<10). Οι έλεγχοι *t* με ποινή μπορούν επίσης να επεκταθούν με διάφορους τρόπους και να εφαρμοστούν και σε πιο γενικές πειραματικές καταστάσεις. Ένα μειονέκτημα των μεθόδων αυτών είναι ότι η μηδενική κατανομή του τροποποιημένου *t* δεν είναι γνωστή. Οι Baldi και Long (Baldi and Long 2001), καθώς και ο Smyth (Smyth, 2005), βασίζονται σε μια τροποποιημένη κατανομή *t* με προσαρμοσμένους βαθμούς ελευθερίας. Από την άλλη πλευρά, μέθοδοι όπως η SAM χρησιμοποιούν μεταθέσεις προκειμένου να υπολογιστεί το False Discovery Rate (FDR) το οποίο θα αναλυθεί εκτενώς στο επόμενο Κεφάλαιο.

3.2.5 Άλλες μέθοδοι

Όπως έχουμε ήδη αναφέρει νωρίτερα, πολλές μελέτες μικροσυστοιχιών εκτιμούν τη διαφορική έκφραση των γονιδίων βασισμένες αποκλειστικά στην αλλαγή διπλώματος. Από την άλλη πλευρά, οι παραλλαγές του ελέγχου *t* αποδίδουν καλύτερα παρέχοντας εκτιμήσεις της στατιστικής σημαντικότητας των γονιδίων. Ωστόσο, ακόμη και αυτοί οι σύγχρονοι στατιστικοί έλεγχοι επιτρέπουν γονίδια με σχετικά μικρή αλλαγή διπλώματος να θεωρούνται στατιστικά σημαντικά κατά εξαιτίας του πολύ μικρού παρονομαστή του εκτιμητή *t*.

Ως εκ τούτου, γίνεται ολοένα και πιο έντονη συζήτηση στη βιβλιογραφία, έτσι ώστε να βρεθούν διαφορικά εκφρασμένα γονίδια τα οποία να πληρούν και τα δύο κριτήρια: του p-value και της αλλαγής διπλώματος. Αρκετοί συγγραφείς επισημαίνουν ότι τα σημαντικά γονίδια είναι αυτά που έχουν ένα αποδεκτό επίπεδο στατιστικής σημαντικότητας και στη συνέχεια, κατατάσσονται ως σημαντικά γονίδια ανάλογα με την τιμή της αλλαγής του διπλώματος (για μια αυθαίρετη τιμή κατωφλίου). Υπάρχουν επίσης συγγραφείς που εφαρμόζουν μια αυθαίρετη τιμή κατωφλίου για την αλλαγή διπλώματος και στη συνέχεια κατατάσσουν τα γονίδια σύμφωνα με την τιμή του p-value. Άλλοι συγγραφείς δηλώνουν γονίδια ως διαφορικά εκφρασμένα αν αυτά έχουν ταυτόχρονα μια τιμή στην αλλαγή διπλώματος

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

μεγαλύτερη από μία δεδομένη τιμή κατωφλίου και ικανοποιείται και το κριτήριο για την τιμή του p-value. Τέτοια συνδυαστικά κριτήρια μπορούν να προσδιορίσουν καλύτερα σύνολα γονιδίων με βιολογική σημασία και να παρέχουν ακόμη καλύτερα αποτελέσματα στις συγκρίσεις γονιδίων εντός της ίδιας πλατφόρμας σε αντίθεση με τη σύγκριση μόνο με την τιμή της αλλαγής του διπλώματος ή της τιμής του p-value (McCarthy and Smyth, 2009).

Η μέθοδος TREAT (έλεγχος *t* σχετικός με ένα κατώφλι) είναι μια επέκταση της εμπειρικής Μπεϋζιανής παραλλαγής του ελέγχου *t* που προτάθηκε από τον Smyth (δηλαδή η μέθοδος limma), και μπορεί να χρησιμοποιηθεί για να δοκιμαστεί εάν η πραγματική διαφορική γονιδιακή έκφραση είναι μεγαλύτερη από μία δεδομένη τιμή κατωφλίου. Στον έλεγχο συμπεριλαμβάνεται μια τιμή κατωφλίου για την αλλαγή διπλώματος και επιτυγχάνονται αξιόπιστες τιμές p-value για τον εντοπισμό γονιδίων με διαφορική έκφραση (McCarthy and Smyth, 2009). Η μέθοδος TREAT έχει αποδειχθεί ότι αποδίδει καλά και σε πραγματικά δεδομένα και σε δεδομένα προσομοίωσης.

Παρόμοιες θεωρήσεις έχουν οδηγήσει στην ανάπτυξη της μεθόδου σταθμισμένης διαφοράς μέσων τιμών (WAD) για την κατάταξη των διαφορικά εκφρασμένων γονιδίων (Kadota et al., 2008). Οι συγγραφείς παρατήρησαν ότι ορισμένα γονίδια τα οποία βρέθηκαν ψευδώς διαφορικά εκφρασμένα τείνουν να εμφανίζουν χαμηλότερα επίπεδα έκφρασης. Με τον τρόπο αυτό, τα πραγματικά διαφορικά εκφρασμένα γονίδια δεν μπορούν να προσδιοριστούν επειδή το σχετικό σφάλμα αυξάνεται σε χαμηλότερες εντάσεις σήματος. Η μέθοδος WAD χρησιμοποιεί τη διαφορά της μέσης τιμής της γονιδιακής έκφρασης και της μέση τιμής της σχετικής έντασης του σήματος κατά τέτοιο τρόπο ώστε διαφορικά εκφρασμένα γονίδια να βρίσκονται πάντα στην κορυφή της κατάταξης για τις διαφορετικές συνθήκες:

$$WAD = \left(\bar{X}_1 - \bar{X}_2\right)\frac{\bar{X} - \min_p\left(\bar{X}\right)}{\max_p\left(\bar{X}\right) - \min_p\left(\bar{X}\right)} \qquad (13)$$

όπου $\bar{X} = \left(\bar{X}_1 + \bar{X}_2\right)/2$ και το max (ή min) δείχνει τη μέγιστη τιμή ή την ελάχιστή τιμή στο διάνυσμά $\bar{X}$ της διαφοράς των μέσω τιμών έκφρασης μεταξύ των p γονίδιων που αναλύθηκαν (σε λογαριθμική κλίμακα). Η μέθοδος WAD συγκρίθηκε με διάφορες άλλες μεθόδους και τα αποτελέσματα έδειξαν ότι υπερτερεί όσον αφορά τόσο την ευαισθησία όσο και την ειδικότητα.

Τέλος, η μέθοδος RankProduct (RP) βασίζεται στον υπολογισμό των γινομένων των βαθμών κατάταξης από την επανάληψη των πειραμάτων, με ένα γρήγορο και απλό τρόπο. Η μέθοδος αυτή αποσκοπεί στην μείωση των προαναφερθέντων προβλημάτων χρησιμοποιώντας την αλλαγή του

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

διπλώματος των γονιδίων και παρέχοντας ταυτόχρονα μια εκτίμηση της στατιστικής σημαντικότητας. Η μέθοδος RP είναι ουσιαστικά μια μη-παραμετρική μέθοδος για την ανίχνευση των διαφορικά εκφρασμένων γονιδίων σε πειράματα μικροσυστοιχιών (Breitling et al., 2004; Breitling and Herzyk, 2005). Τα γονίδια κατατάσσονται σύμφωνα με την τιμή της αλλαγής του διπλώματος και στη συνέχεια η ανάλυση γίνεται ξεχωριστά για τα υπερ-εκφρασμένα και τα υπο-εκφρασμένα γονίδια. Για παράδειγμα, όσον αφορά το υπερ-εκφρασμένο γονίδιο g με i = 1, 2, ... k επαναλήψεις, το γινόμενο των βαθμών θα δίνεται από το γεωμετρικό μέσο:

$$RP_g^{up} = \left(\prod_k r_{g,i}^{up}\right)^{1/k} \quad (14)$$

Η μέθοδος RP είναι διαθέσιμη ως πακέτο στην R (RankProd) και υποστηρίζεται επίσης και διαδικτυακά μέσω του εργαλείου RankProdIt στην ιστοσελίδα (http://strep-microarray.sbs.surrey.ac.uk/RankProducts/).

Στο Κεφάλαιο αυτό παρουσιάστηκε μια ανασκόπηση όλων των μεθοδολογιών που χρησιμοποιούνται για τον προσδιορισμό της στατιστικής σημαντικότητας των γονιδίων από μελέτες μικροσυστοιχιών DNA. Πολλές από αυτές τις μεθόδους θα τις δούμε και στο επόμενο Κεφάλαιο γιατί μπορούν να γενικευτούν και να χρησιμοποιηθούν και ως μέθοδοι μετα-ανάλυσης με σκοπό την εύρεση διαφορικά εκφρασμένων γονιδίων από περισσότερες από μια μελέτες μικροσυστοιχιών που ερευνούν το ίδιο βιολογικό ερώτημα. Η ανασκόπηση των μεθόδων ανάλυσης και η υλοποίηση των μεθόδων ανάλυσης που παρουσιάστηκαν στο παρόν Κεφάλαιο έχει δημοσιευθεί ως Κεφάλαιο στο επιστημονικό βιβλίο Methods in Molecular Biology (Kontou et al., 2016a).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 4

Μεθοδολογία Μετα-Ανάλυσης Δεδομένων Γονιδιακής Έκφρασης από Μικροσυστοιχίες DNA

Η τεχνολογία των μικροσυστοιχιών είναι ένα ευρέως χρησιμοποιούμενο εργαλείο για την ταυτόχρονη ανάλυση της έκφρασης χιλιάδων γονιδίων με αποτέλεσμα τη ραγδαία παραγωγή συνόλων δεδομένων γονιδιακής έκφρασης τα οποία χρειάζονται περαιτέρω ανάλυση. Η μεθοδολογία της μετα-ανάλυσης μπορεί να χρησιμοποιηθεί στην τεχνολογία των μικροσυστοιχιών έτσι ώστε να συνδυαστούν τα αποτελέσματα των επιμέρους μελετών ή πειραμάτων για να εξαχθεί ένα τελικό συμπέρασμα. Συνήθως η μετα-ανάλυση σε δεδομένα γονιδιακής έκφρασης χρησιμοποιείται για την εύρεση γονιδίων τα οποία εκφράζονται διαφορικά (differentially expressed) μεταξύ παθολογικών και φυσιολογικών καταστάσεων. Σήμερα, οι βάσεις δεδομένων GEO (http://www.ncbi.nlm.nih.gov/geo/) και ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) παρέχουν τα κανονικοποιημένα αλλά και τα ανεπεξέργαστα δεδομένα από ένα πλήθος πειραμάτων μικροσυστοιχιών από διάφορους οργανισμούς, επιτρέποντας με τον τρόπο αυτό τη σύγκριση των προφίλ γονιδιακής έκφρασης. Στην παρακάτω ενότητα, παρέχουμε έναν πρακτικό οδηγό για τα βήματα που απαιτούνται προκειμένου να διεξαχθεί σωστά μια μετα-ανάλυση δεδομένων μικροσυστοιχιών.

4.1 Βήματα μετα-ανάλυσης μικροσυστοιχιών

Βήμα 1: Επιλογή των κατάλληλων σετ δεδομένων μικροσυστοιχιών

Το πρώτο και πιο σημαντικό βήμα σε μια πειραματική μελέτη μικροσυστοιχιών και μετέπειτα στη μετα-ανάλυση είναι να αναφέρει σαφώς τους στόχους. Η μετα-ανάλυση επιτρέπει τον εντοπισμό διαφορικά εκφρασμένων γονιδίων μεταξύ πολλαπλών δειγμάτων από διαφορετικές μελέτες, τον εντοπισμό διαφορικά συν-εκφρασμένων γονιδίων και συμβάλλει στη δημιουργία δικτύων γενετικών αλληλεπιδράσεων. Το δεύτερο στάδιο της μετα-ανάλυσης είναι να θέσει τα κριτήρια επιλεξιμότητας των μελετών που θα συμπεριλάβει, είτε βιολογικά (π.χ., τύπος ιστού, ασθένεια) είτε τεχνικά (π.χ., τύπος μικροσυστοιχίας, τεχνολογική πλατφόρμα). Με βάση αυτά τα κριτήρια, πραγματοποιείται αναζήτηση στη βιβλιογραφία με τη χρήση των κατάλληλων όρων με σκοπό να ανακτηθούν οι σχετικές μελέτες. Οι μελέτες αυτές μπορούν να συμπληρωθούν επιπλέον από πειράματα μικροσυστοιχιών που βρίσκονται

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

διαθέσιμα σε δημόσιες βάσεις δεδομένων. Τα δεδομένα των μικροσυστοιχιών πρέπει να ακολουθούν το πρότυπο MIAME (Minimum Information About a Microarray Experiment) έτσι ώστε να μην υπάρχει πλεονάζουσα πληροφορία σχετικά με τα αποτελέσματα του πειράματος μικροσυστοιχιών (Brazma, Hingamp et al. 2001).

Βήμα 2: Ανάκτηση δεδομένων μικροσυστοιχιών από τις μελέτες

Τα γονίδια που βρέθηκαν να εκφράζονται διαφορικά σε μία δεδομένη μελέτη συνιστούν τη δημοσιευμένη λίστα γονίδιων (PGLs) τα οποία είτε περιλαμβάνονται στο κύριο κείμενο της δημοσίευσης είτε παρέχονται ως συμπληρωματικό υλικό. Οι πίνακες δεδομένων γονιδιακής έκφρασης (GEDM) περιέχουν τις τιμές έκφρασης του κάθε ανιχνευτή (probe) του γονιδίου για κάθε δείγμα. Ο πίνακας δεδομένων γονιδιακής έκφρασης δεν μπορεί να χρησιμοποιηθεί άμεσα για μετα-ανάλυση, λόγω των διαφορετικών αλγορίθμων που χρησιμοποιούνται για την επεξεργασία των ακατέργαστων δεδομένων στις αρχικές μελέτες, οι οποίες μπορεί να παράγουν ετερογενή και μη συγκρίσιμα αποτελέσματα.

Βήμα 3: Προεπεξεργασία των συνόλων δεδομένων από διαφορετικές πλατφόρμες μικροσυστοιχιών

Προκειμένου να διεξαχθεί μια σωστή ανάλυση όλων των διαφορετικών συνόλων δεδομένων, πρέπει να εξαλειφθεί η προκατάληψη (bias) που εισήγαγαν οι αλγόριθμοι επεξεργασίας δεδομένων. Για το σκοπό αυτό, θα πρέπει να ανακτηθούν τα αρχικά σετ δεδομένων, όπως τα αρχεία CEL και να μετατραπούν στη μορφή πίνακα δεδομένων γονιδιακής έκφρασης κατάλληλο για μετα-ανάλυση. Διαφορετικές μελέτες από την ίδια πλατφόρμα θα πρέπει να επεξεργαστούν χρησιμοποιώντας τον ίδιο αλγόριθμο. Σε περίπτωση που οι μελέτες που διεξήχθησαν χρησιμοποίησαν διαφορετικές πλατφόρμες, συνιστάται τα δεδομένα να επεξεργάζονται με συγκρίσιμους αλγόριθμους, προκειμένου να μπορούν να συνδυαστούν σωστά στη μετα-ανάλυση και να ελαχιστοποιηθεί η προκατάληψη λόγω της διαφορετικής επεξεργασίας των αρχικών δεδομένων.

Βήμα 4: Αντιστοίχιση ανιχνευτών και γονιδίων

Τα σύνολα δεδομένων χρησιμοποιούν ως σύμβολα/αναγνωριστικά γονιδίων τα ονόματα από τη βάση δεδομένων UniGene ή RefSeq και αναφέρονται συλλογικά ως GeneIDs. Πολλαπλοί ανιχνευτές (probes) μπορούν να υβριδοποιηθούν με το ίδιο GeneID. Για παράδειγμα μια εγγραφή γονιδίου στη UniGene αντιπροσωπεύει ένα σύμπλεγμα αλληλουχιών που αντιστοιχούν σε ένα μοναδικό γονίδιο. Αντίστροφα,

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

υπάρχουν μη ειδικοί ανιχνευτές οι οποίοι μπορούν να υβριδοποιούνται με πολλαπλά GeneIDs λόγω ατελούς εξειδίκευσης. Υπάρχουν επίσης ανιχνευτές με ανεπαρκείς πληροφορίες στην αλληλουχία τους που δεν μπορούν να υβριδοποιηθούν με οποιοδήποτε GeneID. Μία προσέγγιση για την επίλυση του προβλήματος "πολλά-προς-πολλά" στην αντιστοίχιση μεταξύ των ανιχνευτών και των γονιδίων είναι να περιληφθούν στη μετα-ανάλυση μόνο ανιχνευτές που συνδέονται με ένα μόνο γονίδιο και να αποκλειστούν οι ανιχνευτές που συνδέονται με περισσότερα από ένα γονίδια. Ωστόσο, με αυτό τον τρόπο μπορεί να χαθεί σημαντική πληροφορία. Εναλλακτική προσέγγιση είναι είτε η χρήση του μέσου όρου των τιμών έκφρασης των ανιχνευτών που αντιστοιχούν στο ίδιο γονίδιο, είτε η χρήση της μέγιστης τιμής έκφρασης μεταξύ αυτών των ανιχνευτών. Κατά συνέπεια, συνιστάται η εφαρμογή της περιγραφικής στατιστικής προκειμένου να μειωθεί η σχέση "πολλά-προς-πολλά" σε "ένα-προς-ένα» μεταξύ των ανιχνευτών και των μοναδικών γονιδίων για κάθε μελέτη (Consortium et al., 2006; Ramasamy et al., 2008; Zeeberg et al., 2004).

Βήμα 5: Επιλογή μεθόδου μετα-ανάλυσης

Η επιλογή της μεθόδου μετα-ανάλυσης εξαρτάται από τον τύπο των δεδομένων (π.χ., δίτιμα δεδομένα, συνεχή δεδομένα, δεδομένα επιβίωσης). Η παρούσα ανασκόπηση θα επικεντρωθεί στη σύγκριση των μικροσυστοιχιών μεταξύ δυο καταστάσεων στις οποίες στόχος είναι να εντοπιστούν τα γονίδια που εκφράζονται διαφορικά μεταξύ των δύο διαφορετικών καταστάσεων. Σε αυτές τις περιπτώσεις, υπάρχουν τρεις γενικές κατηγορίες στατιστικών μεθόδων μετα-ανάλυσης που χρησιμοποιούν μεγέθη επίδρασης, p-value και βαθμούς κατάταξης.

4.2 Μέθοδοι μετα-ανάλυσης

4.2.1 Μέθοδος βασισμένη στο t-test

Η πρώτη στατιστική μέθοδος που χρησιμοποιείται για μετα-ανάλυση είναι μια τυποποιημένη προσέγγιση με τη χρήση σταθερών ή τυχαίων επιδράσεων. Η μέθοδος *t*-test χρησιμοποιείται ευρέως στη μετα-ανάλυση δεδομένων από μικροσυστοιχίες (Choi et al., 2003). Ως μέγεθος επίδρασης χρησιμοποιείται η τυποποιημένη διάφορα των μέσων τιμών έκφρασης των δειγμάτων ελέγχου και αναφοράς:

$$d_i = \frac{X_{1i} - X_{2i}}{S_{pi}} \qquad (15)$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

όπου X_{1i} και X_{2i} είναι οι μέσες τιμές των δυο υπο μελέτη ομάδων (ελέγχου και αναφοράς) από τη σύγκριση της i μελέτης και S_{pi} είναι η τυπική απόκλιση η οποία δίνεται από τον τύπο:

$$S_{pi} = \sqrt{\frac{(n_{1i}-1)S_{1i}^2 + (n_{1i}-1)S_{1i}^2}{n_{1i}+n_{1i}-2}} \qquad (16)$$

Στη βιβλιογραφία, η εκτίμηση του δείγματος της τυποποιημένης διαφορά μέσων τιμών αναφέρεται ως *d* του Cohen (Cohen, 1988). Παρ' όλα αυτά, το *d* έχει την τάση να υπερεκτιμάται σε μικρά δείγματα. Αυτή η προκατάληψη (bias) που εισήγαγε το *d* μπορεί να διορθωθεί χρησιμοποιώντας έναν άλλο εκτιμητή το λεγόμενο «Hedges' *g*» ο οποίος παράγει μια περισσότερο αμερόληπτη εκτίμηση. Ένας παράγοντας διόρθωσης, που ονομάζεται *J*, χρησιμοποιείται για να μετατρέψει το *d* σε Hedges' *g*. Παρά το γεγονός ότι υπάρχει ένας ακριβής τύπος για τον *J*, οι ερευνητές χρησιμοποιούν συχνά μια προσέγγιση η οποία δίνεται από την εξίσωση: $g_i = Jd_i = d_i - 3d_i/(4n_i - 9)$. Η εκτιμώμενη διακύμανση του *d* δίνεται από τον τύπο:

$$\mathrm{var}(d_i) = s_i^2 = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right) + \frac{d_i^2}{2(n_{1i}+n_{2i})} \qquad (17)$$

Όταν χρησιμοποιείται το *g* τότε ισχύει $\mathrm{var}(g)=J^2\mathrm{var}(d)$. Σε κάθε περίπτωση είναι εύκολο να υπολογιστεί ένας συνολικός εκτιμητής *d* (ή *g*) από τις επιμέρους μελέτες:

$$d = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i} \qquad (18)$$

Αυτή είναι η γνωστή εκτίμηση μετα-ανάλυσης με τη μέθοδο του σταθμίσματος με το αντίστροφο της διακύμανσης που χρησιμοποιείται στη μετα-ανάλυση με $w_i = 1/s_i^2$ (Normand, 1999; Petiti, 1994). Η παραπάνω μέθοδος υποθέτει ομοιογένεια της επίδρασης μεταξύ των μελετών κάτι το οποίο αποτελεί μια αδύναμη υπόθεση, ειδικά στην περίπτωση των μικροσυστοιχιών. Στην περίπτωση ύπαρξης ετερογένειας μεταξύ των μελετών, υποθέτουμε δηλαδή ότι η πραγματική επίδραση ποικίλλει μεταξύ των μελετών, $d_i \sim N\left(d, s_i^2 + \tau^2\right)$ και ως εκ τούτου, πρέπει να υπολογιστεί η μεταξύ των μελετών διακύμανση (τ^2) (μοντέλο τυχαίων επιδράσεων). Η πιο συχνά χρησιμοποιούμενη μέθοδος για την εκτίμηση του τ^2 είναι η μη επαναληπτική μέθοδος που προτείνεται από τους Dersimonian και Laird (DerSimonian and Laird, 1986). Σε περίπτωση που ισχύει $\tau^2 = 0$, τότε οι εκτιμήσεις του μοντέλου τυχαίων επιδράσεων και του μοντέλου σταθερών επιδράσεων συμπίπτουν. Στην περίπτωση του μοντέλου τυχαίων επιδράσεων, τα βάρη υπολογίζονται από τον τύπο:

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

$$w_i = \left(\tau^2 + s_i^2\right)^{-1} \quad (19)$$

και στη συνέχεια η εξίσωση (4) εφαρμόζεται με σκοπό να επιτευχθεί η συνολική εκτίμηση. Σε κάθε περίπτωση, τα συμπεράσματα για το συνολικό αποτέλεσμα πραγματοποιούνται με βάση την κανονική προσέγγιση, δεδομένου ότι:

$$\mathrm{var}\left(d\right) = \frac{1}{\sum_{i=1}^{k} w_i} \quad (20)$$

Όπως έχουμε ήδη επισημάνει, η προσέγγιση αυτή είναι κλασική στη μετα-ανάλυση και έτσι υποστηρίχθηκε από νωρίς στη βιβλιογραφία (Choi et al., 2003; Stevens and Doerge, 2005). Ωστόσο, δεν μπορεί να χειριστεί το πρόβλημα του μικρού μεγέθους του δείγματος και τα μη κανονικά δεδομένα και έτσι οι περισσότεροι συγγραφείς προτείνουν ένα είδος διόρθωσης για τον υπολογισμό της στατιστικής σημαντικότητας. Ως εκ τούτου, αντί η εκτίμηση να βασίζεται στην κανονική προσέγγιση, προτείνουν τη χρήση του ελέγχου μετάθεσης (permutation test) η οποία αναλύθηκε εκτενώς στο Κεφάλαιο 3. Ο Choi και οι συνεργάτες του (Choi et al., 2003) προτείνουν τη μετάθεση για τον υπολογισμό των p-value, ωστόσο, μια ταχύτερη λύση προσφέρεται στο πακέτο *GeneMeta* του Bioconductor της R, η οποία προϋποθέτει μια κανονική κατανομή για τα Z-scores μετά τον έλεγχο της αξιοπιστίας αυτής της υπόθεσης. Σε γενικές γραμμές, όλες οι μέθοδοι με αναδειγματοληψία (Bootstrap) που αναφέρθηκαν στο Κεφάλαιο 3 μπορούν να χρησιμοποιηθούν για την καλύτερη εκτίμηση του τελικού αποτελέσματος, αλλά από όσο γνωρίζουμε τέτοια εφαρμογή δεν έχει γίνει μέχρι στιγμής στη μέτα-ανάλυση. Η μέθοδος Bootstrap ή η μέθοδος με μετάθεση μπορούν επίσης να χρησιμοποιηθούν με διαφορετικούς τρόπους. Μια επιλογή θα ήταν να εκτελεστεί η ανάλυση για κάθε μελέτη ξεχωριστά, να ληφθούν οι διορθωμένες εκτιμήσεις της διακύμανσης και στη συνέχεια να χρησιμοποιηθούν για τον υπολογισμό των βαρών της μετα-ανάλυσης. Μια άλλη επιλογή θα ήταν να εκτελεστεί η ανάλυση σε ένα μόνο βήμα χρησιμοποιώντας τη στρατηγική αναδειγματοληψίας (Bootstrap ή μετάθεση) με ένα στρωματοποιημένο τρόπο, με τον οποίο οι μελέτες αντιμετωπίζονται ως στρώματα στη δειγματοληψία.

Μια άλλη προσέγγιση αποτελεί η χρήση του λόγου των μέσων τιμών μεταξύ των δυο καταστάσεων αντί της τυποποιημένης διαφοράς (Hu et al., 2009). Αυτή η προσέγγιση έχει το πλεονέκτημα ότι χρησιμοποιεί ένα μέτρο που σχετίζεται με τη γνωστή αλλαγή διπλώματος (Fold Change). Έτσι, ο εκτιμητής θα είναι:

$$\gamma_i = \log\left(\frac{X_{1i}}{X_{2i}}\right) \quad (21)$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

με διακύμανση:

$$\mathrm{var}(\gamma_i) = s_i^2 = \frac{1}{n_{1i}} \frac{S_{1i}^2}{X_{1i}^2} + \frac{1}{n_{2i}} \frac{S_{2i}^2}{X_{2i}^2} \qquad (22)$$

Ο λόγος των μέσων έχει επίσης χρησιμοποιηθεί και σε άλλα δεδομένα εκτός από την έκφραση του γονιδίου, και γενικά δίνει καλά αποτελέσματα ακόμα και σε μικρά δείγματα (Friedrich et al., 2008). Οι μέθοδοι αναδειγματοληψίας και μετάθεσης που αναφέρθηκαν παραπάνω εφαρμόζονται επίσης σε αυτό το μέγεθος της επίδρασης.

Οι προαναφερθείσες μέθοδοι μετα-ανάλυσης μπορεί εύκολα να επεκταθούν και σε ένα Μπεϋζιανό μοντέλο (Sutton and Abrams, 2001). Διάφορες μελέτες έχουν πραγματοποιηθεί για το σκοπό αυτό και ο πηγαίος κώδικα για τη δημιουργία του μοντέλου είναι διαθέσιμος (Conlon et al., 2007; Conlon et al., 2006). Σε γενικές γραμμές, ο Conlon και οι συνεργάτες του (Conlon et al., 2007; Conlon et al., 2006) χρησιμοποιούν στα μοντέλα τους μια δομή παρόμοια με εκείνη των Gotardo και συνεργατών που χρησιμοποιείται στην ανάλυσης μιας μόνο μελέτης και την επεκτείνουν προσθέτοντας ένα ακόμη επίπεδο στο μοντέλο τους έτσι ώστε να λειτουργεί σε πολλές μελέτες. Το κυριότερο πρόβλημα με τις Μπεϋζιανές μεθόδους είναι η αυξημένη υπολογιστική πολυπλοκότητα και ο χρόνος που απαιτείται για την εκτέλεση της ανάλυσης. Ειδικά όταν διερευνάται ένας μεγάλος αριθμός γονιδίων περιορίζεται η δυνατότητα εφαρμογής τους. Ο πηγαίος κώδικας WinBUGS που χρησιμοποιείται για την εκτίμηση του μοντέλου τον Conlon και συνεργατών είναι διαθέσιμος στην ιστοσελίδα: http://people.math.umass.edu/~conlon/research/BayesPoolMicro/.

Τέλος, έχει προταθεί και μια άλλη προσέγγιση (βασισμένη στη μέθοδο *limma* που είδαμε στο προηγούμενο Κεφάλαιο) η οποία χρησιμοποιεί τροποποιημένα ή ποινικοποιημένα μεγέθη επίδρασης αντί των τυπικών μεγεθών επίδρασης που χρησιμοποιούνται στη μετα-ανάλυση. Αυτή είναι μια μέθοδος δύο βημάτων. Το πρώτο βήμα στηρίζεται σε μια παραλλαγή του ελέγχου *t* (Marot et al., 2009). Στη συνέχεια, παρατηρώντας ότι $t = d\sqrt{n}$, πραγματοποιείται μετα-ανάλυση με τη χρήση του μοντέλου τυχαίων επιδράσεων. Μια άλλη τροποποίηση της παραπάνω προσέγγισης είναι αντί της χρήσης του *d*, να χρησιμοποιείται ο ακριβής υπολογισμός του Hedges' *g*. Η προσέγγιση αυτή υλοποιείται στο πακέτο *metaMA* (https://cran.r-project.org/web/packages/metaMA/index.html) της R.

Αρκετές μέθοδοι μετα-ανάλυσης για την ανίχνευση των διαφορικά εκφρασμένων γονιδίων (συμπεριλαμβανομένων των μεθόδων συνδυασμού των p-value και του γινομένου των βαθμών κατάταξης που θα παρουσιαστούν στις επόμενες ενότητες) είναι υλοποιημένες σε πακέτα της R, όπως το *GeneMeta* και το *metaMA*. Το πιο ολοκληρωμένο πακέτο είναι το *MetaDE*, το οποίο προσφέρει

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

επίσης μεθόδους για την επεξεργασία των δεδομένων καθώς και για τη γραφική απεικόνιση των αποτελεσμάτων (Wang et al., 2012a). Το στατιστικό λογισμικό Stata στερείται από μια εντολή μετα-ανάλυσης ειδικά για μικροσυστοιχίες, αλλά πολλές από τις μεθόδους που αναφέρονται εδώ μπορούν εύκολα να εφαρμοστούν.

Στην παρούσα διατριβή δημιουργήθηκαν δυο προσεγγίσεις για μετα-ανάλυση δεδομένων μικροσυστοιχιών με τη χρήση μοντέλου τυχαίων επιδράσεων στο στατιστικό λογισμικό Stata οι οποίες αποτελούν επέκταση των μεθόδων ανάλυσης που υλοποιήθηκαν στο Κεφάλαιο 3. Η μία προσέγγιση είναι να εκτελεστεί η ανάλυση σε κάθε μελέτη ξεχωριστά (χρησιμοποιώντας τη μέθοδο Bootstrap ή μετάθεσης) και στη συνέχεια να πραγματοποιηθεί η κλασική μετα-ανάλυση με τη χρήση της εντολής *metan*. Η άλλη προσέγγιση θα είναι να εκτελεστεί μετα-ανάλυση σε ένα μόνο βήμα με ταυτόχρονη εκτέλεση της προσομοίωσης της μεθόδου Bootstrap ή μετάθεσης. Τόσο η μέθοδος Bootstrap όσο και η μέθοδος μετάθεσης εκτελούνται με ένα στρωματοποιημένο τρόπο στον οποίο οι μελέτες αντιμετωπίζονται ως στρώματα. Όλα τα απαραίτητα στοιχεία που χρειάζονται για την εκτέλεση των εντολών αυτών βρίσκονται διαθέσιμα στο ευρύ κοινό μέσω της ιστοσελίδας www.compgen.org/tools/microarrays.

4.2.2 Μέθοδος συνδυασμού των p-value

Μια άλλη κατηγορία μεθόδων μετα-ανάλυσης που χρησιμοποιείται ευρέως σε δεδομένα από μικροσυστοιχίες είναι αυτή του συνδυασμού των p-value. Στις αρχές του 1930, ο Fisher ανέπτυξε την πρώτη μεθοδολογία μετα-ανάλυσης η οποία συνδυάζει τις τιμές των p-value από ανεξάρτητες μελέτες (Fisher, 1946). Ο Fisher πρότεινε ότι αν οι τιμές των *p*-value των *k* ανεξάρτητων δείγματων είναι τυχαίες μεταβλητές που ακολουθούν ομοιόμορφη κατανομή τότε το άθροισμα των λογάριθμων τους θα ακολουθεί την χ^2 κατανομή με $2k$ βαθμούς ελευθερίας:

$$U = -2\sum_{i=1}^{k} \log\left(p_i\right) = -2\log\left(\prod_{i=1}^{k} p_i\right) \qquad (23)$$

Οι Bailey και Gribskov έδειξαν ότι η ίδια πιθανότητα μπορεί να υπολογιστεί γρηγορότερα με τον αλγόριθμο QFAST χωρίς να εξαρτάται από την κατανομή χ^2 (Bailey and Gribskov, 1998). Ο Edgington πρότεινε να χρησιμοποιηθεί το άθροισμα των p-value, προκειμένου να ληφθεί μια συγκεντρωτική εκτίμηση (Edgington, 1972)

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

$$p = \frac{\left(\sum_{i=1}^{k} p_i\right)^k}{k!} \qquad (24)$$

Μια πιο εξελιγμένη μέθοδος παρουσιάστηκε από τον Zaykin και τους συνεργάτες του, η λεγόμενη μέθοδος περικομμένου γινομένου (truncated product method/TPM). Στη μέθοδο αυτή χρησιμοποιούνται μόνο τα γινόμενα των p-value τα οποία είναι μικρότερα από μια συγκεκριμένη τιμή αποκοπής *(τ)*, υπό τη γενική υπόθεση ότι όλες οι *k* υποθέσεις είναι αληθείς (Zaykin et al., 2002). Η εξίσωση που χρησιμοποιείται είναι:

$$W = \prod_{i=1}^{k} (p_i)^{I(p_i \le \tau)} \qquad (25)$$

Οι συγγραφείς παρέχουν και μια πιο αναλυτική εξίσωση για αυτήν την τιμή του p-value:

$$P(W \le w) = \sum_{i=1}^{k} \binom{k}{r} (1-\tau)^{k-r} \left(w \sum_{s=0}^{r-1} \frac{(r\log\tau - \log w)^s}{s!} I\left(w \le \tau^r\right) + \tau^r I\left(w > \tau^r\right) \right) \qquad (26)$$

όπου *r* είναι ο αριθμός των p-value που είναι μικρότερα από την τιμή *τ*. Επιπλέον, ο Zaykin και οι συνεργάτες του έδειξαν από την προσομοίωση ότι η παραπάνω εξίσωση διαθέτει μεγάλη στατιστική ισχύ για την ανίχνευση αποκλίσεων από τη συνολική υπόθεση. Πιο συγκεκριμένα, όταν το *τ* είναι ελάχιστο *(τ = min*), η τιμή του p είναι το αποτελέσμα της γνωστής μεθόδου διόρθωσης Sidak και όταν το τ=1 τότε το *W* γίνεται $W = \prod_{i=1}^{k} p_i$. Η μέθοδος αυτή σε αυτήν την περίπτωση, είναι ταυτόσημη με τη μέθοδο του Fisher, χωρίς την ανάγκη της αναζήτησης της αθροιστικής πιθανότητας στην χ^2 κατανομή:

$$P(W \le w) = w \sum_{i=0}^{k-1} \frac{(-\log w)^i}{i!} \qquad (27)$$

Ο πηγαίος κώδικας για την εφαρμογή της μεθόδου TPM είναι διαθέσιμος στην ιστοσελίδα: http://statgen.ncsu.edu/zaykin/tpm/. Οι διαφορετικές προσεγγίσεις που χρησιμοποιούνται για το συνδυασμό των p-value έχουν συγκριθεί σε διάφορες μελέτες αξιολόγησης (Cousins, 2007; Loughin, 2004). Οι μέθοδοι που παρουσιάστηκαν σε αυτήν την ενότητα εφαρμόζονται στην εντολή *metap* η οποία είναι διαθέσιμη στο Stata και στην R.

Παρ' όλα αυτά, συνδυάζοντας σε μια μετα-ανάλυση τα p-value μπορούν να παρουσιαστούν σοβαρά προβλήματα σε σχέση με το να συνδυάζονται μεγέθη επίδρασης, όπως για παράδειγμα όταν οι διαφορετικές μελέτες ελέγχουν διαφορετικές μηδενικές υποθέσεις. Επιπλέον, στο συνδυασμό των p-value, δεν λαμβάνεται υπόψη η κατεύθυνση της συσχέτισης και ως εκ τούτου όλα τα p-value θα πρέπει

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

να είναι μονόπλευρα, αλλιώς τα υπερ- και τα υπο-εκφραζόμενα γονίδια θα πρέπει να συνδυαστούν ξεχωριστά. Τέλος, όλες αυτές οι μέθοδοι δεν μπορεί να ποσοτικοποιήσουν το μέγεθος της συσχέτισης (το μέγεθος της επίδρασης), και το σημαντικότερο, δεν μπορούν να λάβουν υπόψη τους τη μεταξύ των μελετών ετερογένεια. Μια μέθοδος που αναπτύχθηκε από τον Stouffer ξεπερνά μερικώς αυτούς τους περιορισμούς, συνδυάζοντας τα Z-score αντί των p-value (Stouffer et al., 1951):

$$\bar{Z} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{28}$$

Η μέθοδος αυτή όπως προτάθηκε αρχικά δεν λαμβάνει υπόψη τη διαφορά μεταξύ του μεγέθους των μελετών. Έτσι, μια σταθμισμένη παραλλαγή της παραπάνω εξίσωσης μπορεί να διατυπωθεί από τον παρακάτω τύπο:

$$\bar{Z} = \frac{\sum_{i=1}^{k} \sqrt{w_i} Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \tag{29}$$

με τα βάρη *w* να είναι ανάλογα με την τετραγωνική ρίζα του μεγέθους του δείγματος για κάθε μελέτη:

$$w_i = \sqrt{n_i} \tag{30}$$

Ωστόσο, ακόμα και αυτή η μέθοδος δεν υπολογίζει τη μεταξύ των μελετών μεταβλητότητα και πρόσφατες ενδείξεις από μελέτες γενετικής συσχέτισης (Zhou et al., 2011) υποστηρίζουν ότι η μέθοδος υπολογισμού των βαρών δεν είναι αποτελεσματική. Ο Zhou και οι συνεργάτες του (Zhou et al., 2011) έδειξαν ότι τα βέλτιστα βάρη είναι ανάλογα της ποσότητας $\left(1/n_{1i} + 1/n_{2i}\right)^{-1}$, παρέχοντας με τον τρόπο αυτό τη βάση για τη διεξαγωγή μετα-ανάλυσης τυχαίων επιδράσεων (ακόμη και χωρίς τα πραγματικά μεγέθη επίδρασης). Αξίζει να σημειωθεί ότι, η ιδιαιτερότητα των πειραμάτων μικροσυστοιχιών επιτρέπει στα (μη διαθέσιμα) μεγέθη επίδρασης να εκτιμηθούν με ακρίβεια ως εξής: από το Z- score υπολογίζεται ένα υποθετικό μέγεθος επίδρασης *d**, το οποίο θα αντιστοιχεί στο ίδιο επίπεδο σημαντικότητας:

$$Z_i = d_i^* / se\left(d_i^*\right) \Rightarrow d_i^* = Z_i se\left(d_i^*\right) \tag{31}$$

Το τυπικό σφάλμα αυτού του υποθετικού μεγέθους επίδρασης δίνεται από την εξίσωση (31). Έτσι, ο τύπος για το *d** είναι:

$$d_i^* = Z_i \sqrt{\left(\frac{n_{1i} + n_{2i}}{n_{1i} + n_{2i} - Z_i^2/2}\right)\left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right)} \tag{32}$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Χρησιμοποιώντας αυτό το (υποθετικό) μέγεθος επίδρασης και τη διακύμανσή του, μπορεί να εφαρμοστεί εύκολα η τυπική μέθοδος μετα-ανάλυσης με τη χρήση του μοντέλου τυχαίων επιδράσεων. Αυτή η προσέγγιση απαιτεί μόνο α) το Z-score, το οποίο είτε μπορεί να βρεθεί άμεσα ή υπολογίζεται από το p-value, β) την κατεύθυνση της συσχέτισης και γ) τον αριθμό των δειγμάτων για κάθε κατάσταση. Αυτή η απλή προσέγγιση, κληρονομεί όλες τις επιθυμητές ιδιότητες της μεθόδου Stouffer και, ταυτόχρονα, εκτελεί βέλτιστη στάθμιση, ποσοτικοποιεί τη συσχέτιση και επιτρέπει τη διεξαγωγή μετα-ανάλυσης τυχαίων επιδράσεων, προκειμένου να υπολογιστεί η μεταξύ των μελετών ετερογένεια. Εάν τα αρχικά δεδομένα αναλύονται με κάποια από τις τυποποιημένες μεθόδους, τότε οι εκτιμώμενες τιμές του *d* είναι ακριβείς. Εάν, ωστόσο, χρησιμοποιείται μία παραλλαγή του ελέγχου *t* ή μια μέθοδος αναδειγματοληψίας, τότε αναμένονται ορισμένες αποκλίσεις στην εκτίμηση του *d*. Παρ 'όλα αυτά, το Z-score και η στατιστική σημαντικότητα (p-value) του συνολικού αποτελέσματος υπολογίζονται με ακρίβεια.

4.2.3 Μέθοδος υπολογισμού του γινομένου των βαθμών κατάταξης (rank product)

Μία άλλη κατηγορία μεθόδων για μετα-ανάλυση αποτελείται από μεθόδους οι οποίες συνδυάζουν τους βαθμούς κατάταξης. Υπάρχουν πολλές διαφορετικές προσεγγίσεις, ο κοινός παρονομαστής των οποίων είναι ότι εάν το ίδιο γονίδιο είναι κατ’ επανάληψη στην κορυφή της λίστας μεταξύ των διαφορετικών μελετών, το γονίδιο αυτό μάλλον είναι διαφορικά εκφρασμένο. Η μέθοδος υπολογισμού του γινομένου των βαθμών κατάταξης (rank product) (Breitling et al., 2004) είναι μια μέθοδος μετα-ανάλυσης η οποία χρησιμοποιείται για μετα-ανάλυση δεδομένων γονιδιακής έκφρασης και αναπτύχθηκε από τον Breitling και τους συνεργάτες του. Αρχικά, υπολογίζεται για κάθε γονίδιο η αλλαγή διπλώματος (fold change) η οποία αντιστοιχεί στο λόγο των εκφράσεων μεταξύ των δειγμάτων ελέγχου και αναφοράς. Στη συνέχεια, τα γονίδια κατατάσσονται ανάλογα με την τιμή του λόγου των εκφράσεων από το μεγαλύτερο στο μικρότερο και σε κάθε γονίδιο αντιστοιχεί ένα αριθμός (βαθμός κατάταξης) ο οποίος αντιπροσωπεύει την κατάταξη του γονιδίου στη λίστα. Η παραπάνω διαδικασία γίνεται για όλα τα γονίδια και για όλες τις μελέτες που συμμετέχουν στη μετα-ανάλυση. Στη συνέχεια ως μέγεθος επίδρασης χρησιμοποιείται το γινόμενο των βαθμών κατάταξης το οποίο προκύπτει από τον πολλαπλασιασμό των βαθμών κατάταξης του κάθε γονίδιου από όλες τις μελέτες.

$$RP_g = (\Pi_i \Pi_k r_{gik})^{\frac{1}{k}}$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Το λογισμικό RankProd που υλοποιεί την παραπάνω μέθοδο είναι διαθέσιμο στο ευρύ κοινό στην ιστοσελίδα: https://www.bioconductor.org/packages/release/bioc/html/RankProd.html. Επιπλέον, μια παραλλαγή της μεθόδου RankProd έχει προταθεί, η Rank Sum, στην οποία ως μέγεθος επίδρασης χρησιμοποιείται το άθροισμα των βαθμών κατάταξης αντί του γινομένου.

Μια παραλλαγή της παραπάνω μεθόδου, η METRADISC (Meta-analysis of Rank Discovery Dataset) (Zintzaras and Ioannidis, 2008) έχει προταθεί και είναι πιο γενική σε σχέση με τη RankProd. Στη συγκεκριμένη μέθοδο οι συγγραφείς προτείνουν ως μέγεθος επίδρασης να χρησιμοποιείται ο μέσος όρος των βαθμών κατάταξης των γονιδίων μεταξύ των μελετών και όχι το γινόμενό τους.

$$R^* = \frac{\sum_{i=1}^{k} R_i}{k}$$

Η κατάταξη μπορεί να έχει προέλθει χρησιμοποιώντας οποιαδήποτε μέθοδο (αλλαγή διπλώματος, διαφορά μέσων τιμών, p-value κλπ) ενώ ο συνολικός μέσος όρος μπορεί να είναι σταθμισμένος ή μη. Με αυτές τις προϋποθέσεις, ο συνολικός σταθμισμένος μέσος όρος μοιάζει αρκετά με τις παραδοσιακές μεθόδους μετα-ανάλυσης και η μεταξύ των μελετών ετερογένεια μπορεί επίσης να υπολογιστεί. Η μέθοδος METRADISC είναι υλοποιημένη στην R (http://www.inside-r.org/node/155959) και βρίσκεται επίσης και υλοποιημένη ως ξεχωριστό εργαλείο το οποίο είναι διαθέσιμο στην ιστοσελίδα (http://biomath.med.uth.gr/). Οι μέθοδοι που χρησιμοποιούν την κατάταξη είναι αρκετά ισχυρές και μπορούν να συνδυάσουν τις επιμέρους μελέτες χρησιμοποιώντας διαφορετικές μεθόδους. Ωστόσο, τα στατιστικά συμπεράσματα βασίζονται σε δοκιμές με μετάθεση Μόντε Κάρλο οι οποίες είναι αρκετά χρονοβόρες.

Οι μέθοδοι που βασίζονται στην κατάταξη προσφέρουν διάφορα πλεονεκτήματα σε σύγκριση με τις παραδοσιακές προσεγγίσεις, συμπεριλαμβανομένου του κριτηρίου FC, λιγότερες παραδοχές σύμφωνα με το μοντέλο και έχουν ευρωστία σε δεδομένα με θόρυβο και με μικρό αριθμό αντιγράφων. Αυτές οι μέθοδοι μπορούν να ξεπεράσουν την ετερογένεια μεταξύ των διαφορετικών συνόλων δεδομένων και να τα συνδυάσουν ώστε να επιτευχθεί αυξημένη ευαισθησία και αξιοπιστία στο τελικό αποτέλεσμα. Το πιο σημαντικό πλεονέκτημα αυτών των μεθόδων είναι ότι δεν απαιτούν την κανονικοποίηση των διαφορετικών συνόλων δεδομένων, χρησιμοποιώντας την ίδια τεχνική και με αυτό τον τρόπο επιλύεται το βασικό ζήτημα της επεξεργασίας των δεδομένων μικροσυστοιχιών πριν την πραγματοποίηση της μετα-ανάλυσης. Επιπλέον, οι μέθοδοι που βασίζονται στην κατάταξη μετασχηματίζουν τις τιμές έκφρασης σε βαθμούς κατάταξης και ως εκ τούτου μπορούν να

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ενσωματωθούν σύνολα δεδομένων που προέρχονται από διαφορετικές πλατφόρμες (μικροσυστοιχίες ολιγονουκλεοτίδιων Affymetrix, δικαναλικές μικοσυστοιχίες cDNA κ.λπ.). Τέλος, οι μέθοδοι που βασίζονται στην κατάταξη είναι αρκετά γενικές και ως εκ τούτου μπορεί να εφαρμοστούν σε διαφορετικούς τύπους δεδομένων, όπως στην πρωτεομική ή στα δεδομένα γενετικής συσχέτισης.

4.3 Μέθοδοι διόρθωσης πολλαπλών συγκρίσεων

Ένα πείραμα μικροσυστοιχιών υπολογίζει την ταυτόχρονη έκφραση χιλιάδων γονιδίων κάτω από συγκεκριμένες συνθήκες. Έτσι, για τα δεδομένα των μικροσυστοιχιών απαιτείται μια στατιστική επιλογή ούτως ώστε να εντοπιστούν τα γονίδια τα οποία εκφράζονται διαφορικά (differentially expressed) και είναι στατιστικώς σημαντικά σε σύγκριση με το σύνολο των γονιδίων τα οποία εξετάζονται στη μικροσυστοιχία. Ο έλεγχος *t* (παραμετρικός έλεγχος) χρησιμοποιείται κυρίως για την ταυτόχρονη σύγκριση δύο συνόλων καθώς συγκρίνει τους μέσους όρους των δυο συνόλων και δείχνει αν υπάρχει μεταξύ τους στατιστικά σημαντική διαφορά (Ramasamy et al., 2008). Οι περισσότεροι έλεγχοι δίνουν ένα συγκεκριμένο p-value για κάθε γονίδιο με το οποίο εκφράζεται αν υπάρχουν στατιστικώς σημαντικές διαφορές στην έκφρασή του.

Όταν ένας έλεγχος *t* εκτελείται, η μηδενική υπόθεση (H_0) είναι ότι δεν υπάρχει διαφορά μεταξύ των επιπέδων έκφρασης του γονιδίου μεταξύ των δυο καταστάσεων, ενώ η εναλλακτική υπόθεση (H_1) είναι ότι τα επίπεδα έκφρασης διαφέρουν. Εάν η τιμή του p-value είναι μικρότερη από το επιλεγμένο επίπεδο σημαντικότητας, τότε η μηδενική υπόθεση απορρίπτεται. Αν υποθέσουμε ότι ισχύει η μηδενική υπόθεση και πραγματοποιηθεί έλεγχος σε 10.000 γονίδια σε επίπεδο σημαντικότητας 5%, τότε 500 γονίδια αναμένεται να βρεθούν στατιστικά σημαντικά απλά και μόνο λόγω τύχης. Στα πειράματα των μικροσυστοιχιών ο αριθμός των γονιδίων που εξετάζονται είναι πολύ μεγάλος και γι' αυτό απαιτείται η χρήση μεθόδων διόρθωσης του p-value καθώς τιμές $p<0.05$ και $p<0.01$ θα δώσουν σίγουρα γονίδια τα οποία είναι λανθασμένα στατιστικά σημαντικά.

Οι μέθοδοι διόρθωσης πολλαπλών συγκρίσεων λαμβάνουν ως είσοδο μια λίστα με p-value τα οποία έχουν προκύψει από τους ελέγχους και επιστρέφουν τη διορθωμένη τιμή p-value. Αυτές οι μέθοδοι χωρίζονται σε δύο κατηγορίες, αυτές που ελέγχουν το ποσοστό σφάλματος FWER (family-wise error rate) και αυτά που ελέγχουν το FDR (False Discovery Rate). Σε γενικές γραμμές, η μέθοδος FWER παρέχει ένα διορθωμένο p-value για ένα σύνολο στο οποίο ισχύει η μηδενική υπόθεση. Συνήθως, αυτό το επίπεδο σημαντικότητας είναι μικρότερο από το μη διορθωμένο. Η πιο γνωστή

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

μέθοδος για τον έλεγχο FWER είναι η διόρθωση Bonferroni (Dudoit, 2000), στην οποία το επίπεδο σημαντικότητας α για κάθε έλεγχο υπολογίζεται από τη διαίρεση του FWER (συνήθως 0,05) με τον αριθμό των ελέγχων (δηλαδή το πλήθος των γονιδίων στην περίπτωση των μικροσυστοιχιών). Έτσι, για 10.000 γονίδια το επίπεδο σημαντικότητας για έναν έλεγχο θα είναι $\alpha = 0{,}05 / 10000 = 5 * 10^{-6}$. Τα γονίδια με p-value $< 5*10^{-6}$ εκφράζονται διαφορικά. Το διορθωμένο p-value σύμφωνα με τη διόρθωση Bonferroni δίνεται από τον τύπο:

$$p_{cor(i)} = p_{(i)} / n \qquad (33)$$

Η διόρθωση Bonferroni εφαρμόζεται εύκολα και διαισθητικά, αλλά είναι πολύ συντηρητική. Μια άλλη συχνά χρησιμοποιούμενη μέθοδος είναι η μέθοδος του Sidak (Sidak, 1967):

$$p_{cor(i)} = 1-(1-p_{(i)})^{1/n} \qquad (34)$$

Άλλες δημοφιλείς μέθοδοι που χρησιμοποιούνται για πολλαπλές διορθώσεις στην ανάλυση μικροσυστοιχιών και στη μετα-ανάλυση είναι οι μέθοδοι που προτείνονται από τον Holland και τους συνεργάτες του (Holland and Copenhaver, 1987) και τον Holm (Holm, 1979).

Οι Benjamini και Hochberg (Benjamini and Hochberg, 1995) πρότειναν μια διαφορετική μέθοδο η οποία ελέγχει το FDR, αντί του FWER. Οι μέθοδοι ελέγχου του FDR έχουν σχεδιαστεί για τον έλεγχο του αναμενόμενου ποσοστού απόρριψης της μηδενικής υπόθεσης που όμως είναι εσφαλμένες απορρίψεις (ψευδώς θετικά). Οι διαδικασίες ελέγχου του FDR παρέχουν λιγότερο αυστηρό έλεγχο των σφαλμάτων τύπου I σε σύγκριση με τις διαδικασίες ελέγχου του FWER. Έτσι, οι διαδικασίες ελέγχου του FDR έχουν μεγαλύτερη στατιστική ισχύ (δηλαδή ανιχνεύουν περισσότερες στατιστικά σημαντικές διαφορές μέσων τιμών έκφρασης), με το κόστος της αύξησης των ποσοστών των σφαλμάτων τύπου I (το οποίο όμως, το κρατούν σε επίπεδο που μπορεί να ελεγχθεί). Στη μέθοδο αυτή, τα p-value κάθε γονιδίου κατατάσσονται από το μικρότερο στο μεγαλύτερο. Με τον τρόπο αυτό, το μικρότερο p-value έχει ένα βαθμό κατάταξης $i = 1$, τότε το επόμενο μικρότερο διαθέτει βαθμό κατάταξης $i = 2$, κ.λπ. Στη συνέχεια, κάθε p-value συγκρίνεται με την τιμή του επιπέδου σημαντικότητας του Benjamini-Hochberg, $(i / n) Q$, όπου i είναι ο βαθμός κατάταξης, n είναι ο συνολικός αριθμός των ελέγχων (γονιδίων), και το Q είναι το επιλεγμένο FDR. Τα p-value τα οποία δείχνουν στατιστικά σημαντική συσχέτιση είναι αυτά για τα οποία ισχύει: $p < (i / n) Q$. Το διορθωμένο p-value (πολλές φορές ονομάζεται q-value) δίνεται από τον τύπο:

$$p_{cor(i)} = \frac{i}{n} p_{(i)} \qquad (35)$$

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Άλλες δημοφιλείς μέθοδοι διόρθωσης πολλαπλών συγκρίσεων που ελέγχουν το FDR στην ανάλυση μικροσυστοιχιών και στη μετα-ανάλυση είναι οι μέθοδοι που προτείνονται από τους Benjamini και Yekutieli (Benjamini and Yekutieli, 2001), Benjamini και Liu (Benjamini and Liu, 1999a, b), Benjamini, Krieger και Yekutieli (Benjamini et al., 2006). Οι μέθοδοι που περιγράφονται παραπάνω υλοποιούνται στην εντολή *multproc* στο Stata και στο πακέτο *multcomp* της R.

4.4 Μετα-ανάλυση δεδομένων γονιδιακής έκφρασης για την εύρεση γονιδίων που εκφράζονται διαφορικά στην εμφάνιση του εμφράγματος του μυοκαρδίου.

4.4.1 Εισαγωγή

Η αθηροσκληρωτική καρδιακή νόσος εκδηλώνεται με την αθηροσκλήρωση και έχει ένα ευρύ φάσμα παθολογικών φαινότυπων. Περιλαμβάνει, μεταξύ άλλων, την ισχαιμική καρδιακή νόσο (IHD), τη στεφανιαία νόσο (CAD), το εγκεφαλικό επεισόδιο και το έμφραγμα του μυοκαρδίου (MI) κοινώς γνωστό ως καρδιακή προσβολή. Οι αθηροσκληρωτικές καρδιακές παθήσεις αντιπροσωπεύουν το κύριο αίτιο νοσηρότητας και θνησιμότητας παγκοσμίως, αντιπροσωπεύοντας 17.300.000 θανάτους ετησίως, με αποτέλεσμα περίπου το ένα τρίτο όλων των θανάτων παγκοσμίως (Mortality and Causes of Death, 2015; Wong, 2014). Η στεφανιαία νόσος είναι μια ομάδα ασθενειών, η οποία συμπεριλαμβάνει τη στηθάγχη, το οξύ στεφανιαίο σύνδρομο, τον αιφνίδιο καρδιακό θάνατο και το έμφραγμα του μυοκαρδίου (Wong, 2014). Το έμφραγμα του μυοκαρδίου είναι σύνθετη και πολυπαραγοντική ασθένεια που οφείλεται στην αλληλεπίδραση γενετικών και περιβαλλοντικών παραγόντων (McPherson and Tybjaerg-Hansen, 2016; Pedersen et al., 2016).

Ευρυγονιδιωματικές μελέτες έχουν αποκαλύψει ένα μεγάλο αριθμό πολυμορφισμών ενός νουκλεοτιδίου (SNPs) που σχετίζονται με το έμφραγμα του μυοκαρδίου. Έτσι, τα αποτελέσματα του γενετικού κινδύνου μπορούν να χρησιμοποιηθούν παράλληλα με τις παραδοσιακές μεθόδους (βαθμολογία Framingham) για τον υπολογισμό του καρδιαγγειακού κινδύνου. Πολλές μελέτες γονιδιακής έκφρασης από μικροσυστοιχίες έχουν ήδη χρησιμοποιηθεί για την εύρεση γονιδίων που μπορούν να χρησιμοποιηθούν ως βιοδείκτες σε διάφορες ασθένειες όπως προκαρκινικές καταστάσεις ασθένειας (Dhanasekaran et al., 2001) ή αυξημένη οξείδωση και φλεγμονή σε ασθενείς με δρεπανοκυτταρική νόσο (Jison et al., 2004). Σήμερα, υπάρχει ένα αυξανόμενο ενδιαφέρον και γίνονται προσπάθειες για τη δημιουργία προφίλ έκφρασης βασισμένων κυρίως στις μικροσυστοιχίες (transcriptomics) για τη διάγνωση της στεφανιαίας νόσου και του εμφράγματος του μυοκαρδίου, καθώς

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

και για την πρόβλεψη του κινδύνου εμφάνισης του εμφράγματος του μυοκαρδίου και του καρδιαγγειακού θανάτου (Kessler et al., 2013; Kim et al., 2014).

Στόχος αυτής της μελέτης ήταν να συγκεντρώσει όλα τα διαθέσιμα δεδομένα γονιδιακής έκφρασης σχετικά με γονίδια που εκφράζονται διαφορικά στο έμφραγμα του μυοκαρδίου και να δημιουργηθεί μετέπειτα γονιδιακή υπογραφή κατάλληλη για την πρόγνωση και τη διάγνωση της ασθένειας. Για το σκοπό αυτό, τα δεδομένα γονιδιακής έκφρασης από μελέτες ασθενών-μαρτύρων ανακτήθηκαν από ανεξάρτητες μελέτες μικροσυστοιχιών και πραγματοποιήθηκε μετα-ανάλυση. Η εφαρμογή της μεθόδου της μετα-ανάλυσης και η περαιτέρω ανάλυση που πραγματοποιήθηκε με χρήση εργαλείων βιοπληροφορικής, οδήγησε στην επιβεβαίωση μοριακών μονοπατιών και δικτύων γονιδίων που συμμετέχουν στις καρδιαγγειακές παθήσεις, αλλά και στην αποκάλυψη νέων γονίδιων που συσχετίζονται με το έμφραγμα του μυοκαρδίου και νέων μονοπατιών που οδηγούν στην ανάπτυξη του μυοκαρδίου και του καρδιαγγειακού θανάτου.

4.4.2 Μέθοδοι

Συλλογή Δεδομένων

Προκειμένου να προσδιοριστούν τα δεδομένα γονιδιακής έκφρασης για το έμφραγμα του μυοκαρδίου, πραγματοποιήθηκε μια ολοκληρωμένη βιβλιογραφική έρευνα στη βάση δεδομένων PubMed (McEntyre and Lipman, 2001) και στη βάση δεδομένων μικροσυστοιχιών GEO (Barrett and Edgar, 2006). Οι λέξεις-κλειδιά που χρησιμοποιήθηκαν για την αναζήτηση στην Pubmed ήταν "microarray" AND ("myocardial ischaemia" OR "myocardial infarction"), ενώ για την GEO η λέξη-κλειδί ήταν "myocardial infarction". Επιλέχθηκαν μελέτες ασθενών-μαρτύρων οι οποίες μετρούν τη γονιδιακή έκφραση στο αίμα σε υγιείς και ασθενείς με έμφραγμα του μυοκαρδίου. Σύνολα δεδομένων που μετρούν την έκφραση του γονιδίου σε ιστούς άλλους από το αίμα, καθώς επίσης και σύνολα δεδομένων που μετρούν την επίδραση των φαρμάκων στις προαναφερόμενες ασθένειες εξαιρέθηκαν από την ανάλυσή μας. Τα επιλεγμένα άρθρα μελετήθηκαν πλήρως και κατεγράφησαν οι βασικές τους πληροφορίες όπως ο κωδικός της μελέτης, το όνομα της εργασίας, η μέθοδος ανάλυσης και ο αριθμός των προς μελέτη γονιδίων. Τα σύνολα δεδομένων, δηλαδή οι πίνακες γονιδιακής έκφρασης ανακτήθηκαν από τη βάση δεδομένων GEO. Δεδομένου ότι στα πειράματα μικροσυστοιχιών πολλαπλοί ανιχνευτές μπορεί να αντιστοιχούν στο ίδιο γονίδιο, πριν από τη διεξαγωγή μετα-ανάλυσης μετατράπηκαν οι «όροι» των ανιχνευτών σε «όρους» γονιδίων. Αυτό έγινε προκειμένου να συνδυαστούν μελέτες από διαφορετικές πλατφόρμες και να επιλυθεί το πρόβλημα "πολλά-προς-πολλά" στην αντιστοίχιση μεταξύ των

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ανιχνευτών και των γονιδίων (Ramasamy et al., 2008). Οι πληροφορίες σχετικά με την αντιστοίχιση των ονομάτων των ανιχνευτών με τα αντίστοιχα ονόματα των γονιδίων αντλήθηκαν από το αρχείο GPL το οποίο είναι συνοδευτικό του αρχείου που περιέχει τον πίνακα γονιδιακής έκφρασης. Στη συνέχεια προκειμένου να διεξαχθεί σωστά η μετα-ανάλυση επιλέχθηκαν μόνο τα γονίδια που υπήρχαν τουλάχιστον σε δύο μελέτες.

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

Εικόνα 4.1. Παρουσίαση ενός τμήματος του πίνακα γονιδιακής έκφρασης, από τη μελέτη με κωδικό GSE66360. Εμφανίζονται πολλαπλοί ανιχνευτές του ίδιου γονιδίου με διαφορετικές τιμές έκφρασης.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

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

Εικόνα 4.2. Παρουσίαση ενός τμήματος του πίνακα γονιδιακής έκφρασης, από τη μελέτη με κωδικό GSE66360. Η τιμή έκφρασης κάθε γονιδίου υπολογίστηκε από τον μέσο όρο των πολλαπλών ανιχνευτών.

Στατιστική ανάλυση

Το *t*-test χρησιμοποιήθηκε για να προσδιοριστούν τα διαφορικά εκφρασμένα γονίδια μεταξύ των δύο ομάδων (ασθενείς-μάρτυρες). Ένα μειονέκτημα του *t*-test για την ανάλυση των δεδομένων των μικροσυστοιχιών είναι ότι σε περίπτωση που τα περισσότερα πειράματα περιέχουν μικρό αριθμό δειγμάτων σε κάθε ομάδα η υπόθεση της κανονικότητας δεν ισχύει. Για να επιλυθεί αυτό, χρησιμοποιήθηκε η μέθοδος Bootstrap (Efron, 1982; Efron and Tibshirani, 1993), η οποία είναι μια στατιστική μέθοδος για την εκτίμηση της κατανομής δειγματοληψίας του εκτιμητή με δειγματοληψία με αντικατάσταση από το αρχικό δείγμα. Όπως αναφέρθηκε παραπάνω, το Bootstrap παρέχει μια ιδανική εναλλακτική μέθοδο όταν δεν υπάρχει ο κατάλληλος τύπος για την κατανομή δειγματοληψίας ή όταν υπάρχουν διαθέσιμοι τύποι αλλά κάνουν ακατάλληλες υποθέσεις (π.χ. μικρό μέγεθος του δείγματος, μη-κανονική κατανομή). Χρησιμοποιήθηκε η μέθοδος Bootstrap με 1000 επαναλήψεις, ώστε να προκύψουν ακριβείς εκτιμήσεις των τυπικών σφαλμάτων.

Τα τυπικά σφάλματα που προέκυψαν από τη μέθοδο Bootstrap, χρησιμοποιήθηκαν στην τυπική διαδικασία μετα-ανάλυσης τυχαίων επιδράσεων με βάση την τυποποιημένη διαφορά μέσων τιμών (Campain and Yang, 2010; Choi et al., 2003). Επιπλέον, εφαρμόστηκαν διάφορες μέθοδοι διόρθωσης

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

πολλαπλών συγκρίσεων προκειμένου να περιοριστούν τα ψευδώς στατιστικά σημαντικά γονίδια. Αυτές οι μέθοδοι χωρίζονται σε δύο κατηγορίες, αυτές που ελέγχουν το Family-Wise Error Rate (FWER) και αυτές που ελέγχουν το False Discovery Rate (FDR). Η πιο κοινή διαδικασία για τον έλεγχο της FWER είναι η διόρθωση Bonferroni (Dudoit, 2000), ενώ άλλες δημοφιλείς μέθοδοι που χρησιμοποιούνται για διόρθωση πολλαπλών συγκρίσεων είναι οι μέθοδοι που προτείνονται από τους Sidak (Sidak, 1967), Holland και συνεργάτες (Holland and Copenhaver, 1987) και τον Holm (Holm, 1979). Οι Benjamini και Hochberg (Benjamini and Hochberg, 1995) πρότειναν μια μέθοδο που ελέγχει το FDR. Η μέθοδος FDR έχει μεγαλύτερη δύναμη (δηλαδή ανιχνεύουν περισσότερες διαφορές ως στατιστικά σημαντικές), με το κόστος της αύξησης των ποσοστών των σφαλμάτων τύπου I. Στην ανάλυση, γονίδια με τιμή p ή q μικρότερη ή ίση με 0,01, για το FWER και FDR αντίστοιχα, θεωρήθηκαν ως στατιστικά σημαντικά. Για όλες τις αναλύσεις χρησιμοποιήθηκε το στατιστικό λογισμικό Stata 13 (StataCorp, 2013).

Βιοπληροφορική ανάλυση

Τα διαφορικά εκφρασμένα γονίδια που βρέθηκαν από τη μετα-ανάλυση, υποβλήθηκαν στο εργαλείο βιοπληροφορικής ανάλυσης STRING v10 (Szklarczyk et al., 2015) για την υπολογιστική ανάλυση των αλληλεπιδράσεων πρωτεϊνών-πρωτεϊνών. Η STRING (Search Tool for the Retrieval of INteracting Genes/proteins) (Szklarczyk et al., 2015) είναι μια ολοκληρωμένη βάση δεδομένων των γνωστών άμεσων και έμμεσων αλληλεπιδράσεων μεταξύ γονιδίων/πρωτεϊνών, που προέρχεται από διάφορες πηγές, όπως βιοχημικές, γενετικές, πειραμάτων βιοφυσικής, αναλύσεις συν-έκφρασης και άλλα. Επιπλέον, με τη χρήση του εργαλείου WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) (Wang et al., 2013) βρέθηκαν στατιστικά σημαντικά βιοχημικά μονοπάτια KEGG (Kanehisa et al., 2016) σε όριο σημαντικότητας FDR ίσο ε 0,001. Τέλος, χρησιμοποιήθηκε μια υψηλής απόδοσης πλατφόρμα ανάλυσης δεδομένων, η bioCompendium (http://biocompendium.embl.de/), που δέχεται ως είσοδο λίστα γονιδίων ή πρωτεϊνών. Συγκρίνει τα αποτελέσματα από διαφορετικές πειραματικές συνθήκες και βάσεις δεδομένων και βρίσκει τις μοριακές λειτουργίες και τις βιολογικές διαδικασίες στις οποίες συμμετέχουν τα γονίδια που δόθηκαν ως είσοδο.

4.4.3 Αποτελέσματα

Αρχικά, από τη βάση δεδομένων Pubmed (https://www.ncbi.nlm.nih.gov/pubmed) ανακτήθηκαν 162 άρθρα και 174 πειράματα μικροσυστοιχιών από τη βάση μικροσυστοιχιών GEO

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

(https://www.ncbi.nlm.nih.gov/geo/) (Barrett and Edgar, 2006). Όλα τα 162 άρθρα από την Pubmed είτε δεν πληρούσαν τα κριτήρια επιλεξιμότητας είτε δεν περιείχαν τα δεδομένα γονιδιακής έκφρασης και έτσι αποκλείστηκαν από τη μετα-ανάλυση. Από τα σύνολα δεδομένων της GEO μόνο τέσσερα πληρούσαν τα κριτήρια επιλεξιμότητας και συμπεριλήφθηκαν στη μετα-ανάλυση. Αυτά περιείχαν δεδομένα για 31180 γονίδια, σε 93 ασθενείς με έμφραγμα του μυοκαρδίου και 89 υγιή άτομα.

Πίνακας 4.1. Τα χαρακτηριστικά των μελετών που χρησιμοποιήθηκαν στη μετα-ανάλυση.

PMID	GEO Dataset	Πλατφόρμα	Ασθενείς	Υγιείς	Αριθμός Ανιχνευτών	Αριθμός Γονιδίων
24801707	GSE48060	Affymetrix Human Genome U133 Plus 2.0 Array	30	22	42450	21037
26025919	GSE60993	Illumina HumanWG-6 v3.0 expression beadchip	7	7	35966	25162
26025918	GSE61144	Sentrix Human-6 v2 Expression BeadChip	7	10	30535	24778
-	GSE66360	Affymetrix Human Genome U133 Plus 2.0 Array	49	50	42450	21037

Από τη μετα-ανάλυση δημιουργήθηκε ένα αρχείο που δίνει για κάθε γονίδιο, την τιμή *t*, την τιμή του τυπικού σφάλματος (standard error) και την τιμή του p-value. Τα γονίδια με p-value<0.05 θεωρήθηκαν στατιστικά σημαντικά, όμως ο μεγάλος αυτός αριθμός γονιδίων φανερώνει την ύπαρξη ψευδώς θετικών αποτελεσμάτων όπως αναφέρθηκε και στην παραπάνω ενότητα και υπάρχει ανάγκη διόρθωσης του p-value με τις μεθόδους διόρθωσης πολλαπλών συγκρίσεων. Όσον αφορά τα αποτελέσματα της μετα-ανάλυσης, η μέθοδος FDR με όριο ίσο με 0,01 (Benjamini and Hochberg, 1995), εντόπισε συνολικά 626 διαφορικά εκφρασμένα γονίδια σε ασθενείς με έμφραγμα του μυοκαρδίου σε σύγκριση με τα υγιή άτομα. Οι άλλες μέθοδοι διόρθωσης πολλαπλών συγκρίσεων Sidak, Bonferroni, Holm και Holland εντόπισαν λιγότερα γονίδια ως στατιστικά σημαντικά. Είναι αξιοσημείωτο ότι οι επιμέρους μελέτες, αν αντιμετωπιστούν απομονωμένα, ανιχνεύουν έναν πολύ μικρό αριθμό γονιδίων που να συσχετίζεται με το έμφραγμα και να έχει βρεθεί και στη μετα-ανάλυση. Αυτό επιβεβαιώνει και την αναγκαιότητα της μετα-ανάλυσης μιας και ανιχνεύει στατιστικά σημαντικά γονίδια από ένα σύνολο μελετών τα οποία δε θα έβγαιναν σημαντικά στις επιμέρους μελέτες. Μια σύγκριση μεταξύ των στατιστικά σημαντικών γονιδίων που βρέθηκαν από κάθε μελέτη ξεχωριστά και

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αυτών που βρέθηκαν από τη μετα-ανάλυση παρουσιάζεται τόσο στο διάγραμμα Venn (εικόνα 4.3) όσο και στον Πίνακα 4.2. Μόνο δυο γονίδια (VNN2 και NCALD) βρέθηκαν κοινά στις 3 επιμέρους μελέτες και στη μετα-ανάλυση.

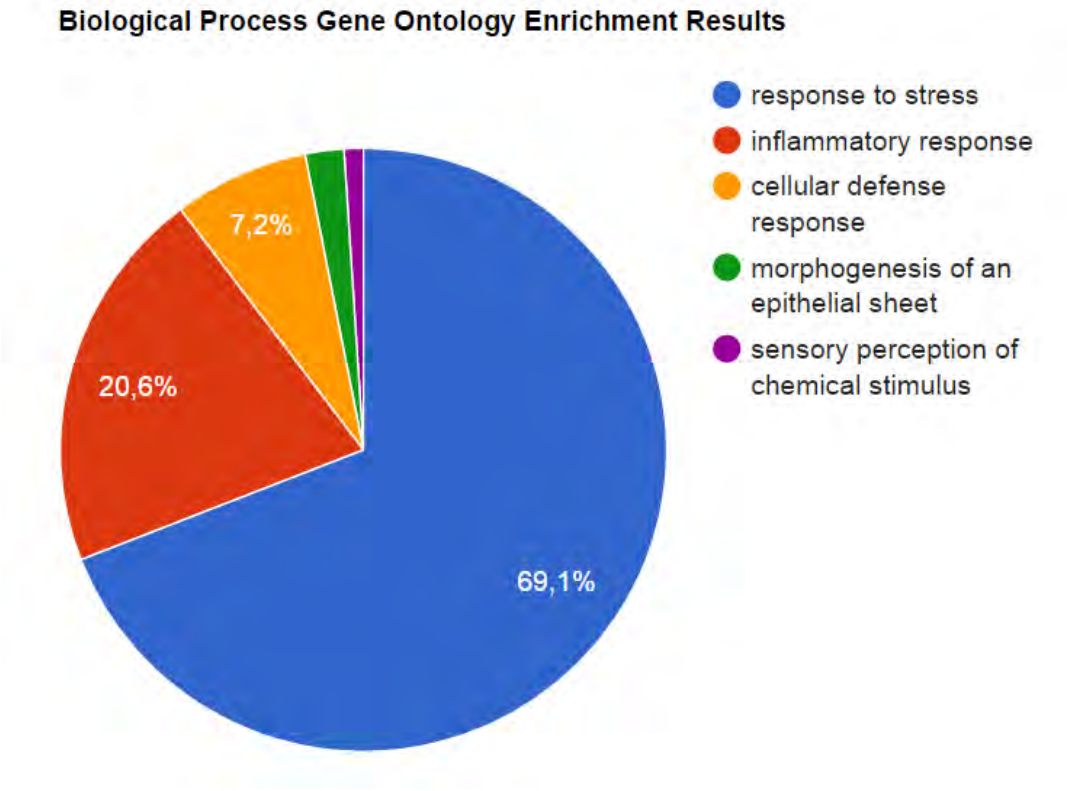
Πίνακας 4.2. Ο αριθμός των στατιστικά σημαντικών γονιδίων από κάθε μελέτη και από τη μετα-ανάλυση ανάλογα με τη μέθοδο διόρθωσης.

	pvalue<0,05	FDR<0,01	Bonferroni	Holm	Sidak	Holland
Μελέτη 1: GSE48060	2845	3	3	3	3	3
Μελέτη 2: GSE60993	4250	112	15	15	15	15
Μελέτη 3: GSE61144	5916	2198	260	263	263	268
Μελέτη 4: GSE66360	4694	1170	325	325	325	325
Μετα-ανάλυση	4306	626	158	158	160	160

Εικόνα 4.3. Διάγραμμα Venn των στατιστικά σημαντικών γονιδίων σε κάθε μελέτη και στη μετα-ανάλυση.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Η περαιτέρω ανάλυση εστίασε στα 626 διαφορικά εκφρασμένα γονίδια στο έμφραγμα του μυοκαρδίου προκειμένου να βρεθεί ο τρόπος με τον οποίο εμπλέκονται στην ασθένεια. Αρχικά, μέσω του εργαλείου Biocompendium ανιχνεύθηκαν οι βιολογικές διαδικασίες στις οποίες συμμετέχουν τα γονίδια αυτά. Είναι αξιοσημείωτο ότι πάνω από το 69% των γονιδίων σχετίζεται με το στρες και το 20,6% των γονιδίων συμμετέχει σε φλεγμονώδεις αποκρίσεις (Εικόνα 4.4).



Εικόνα 4.4. Βιολογικές διαδικάσίες στις οποίες συμμετέχουν τα διαφορικά εκφρασμένα γονίδια στο έμφραγμα του μυοκαρδίου.

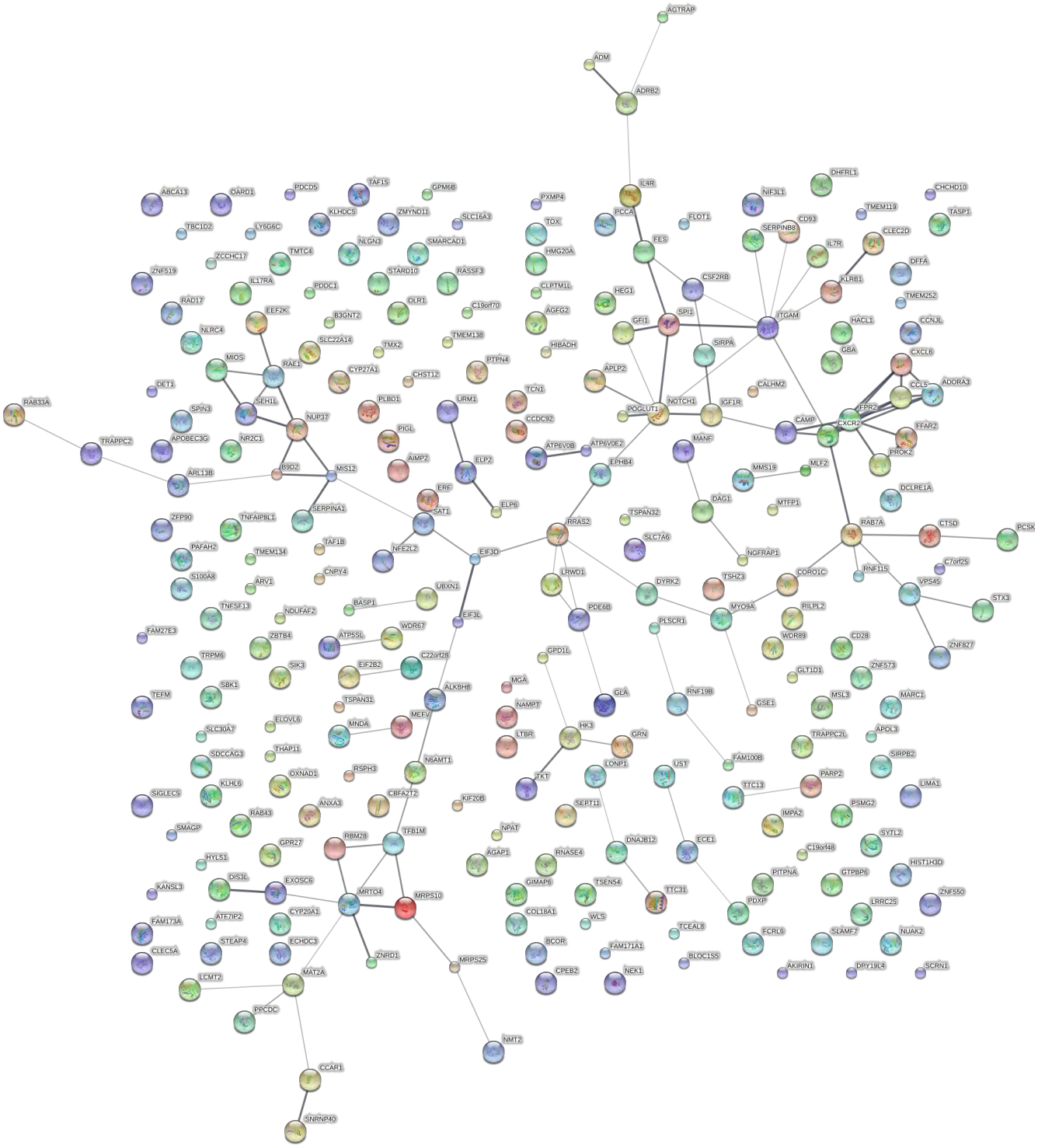
Χρησιμοποιώντας το εργαλείο ανάλυσης STRING καταφέραμε να απεικονίσουμε όλες τις αλληλεπιδράσεις των διαφορικά εκφρασμένων γονιδίων. Στο δίκτυο των πρωτεϊνικών αλληλεπιδράσεων υπάρχουν πολλά γονίδια τα οποία δεν αλληλεπιδρούν με άλλα, υπάρχουν όμως και γονίδια τα οποία σχηματίζουν ομάδες, κάτι το οποίο δείχνει ότι τα γονίδια αυτά εμπλέκονται στις ίδιες λειτουργίες (Εικόνα 4.5). Εντοπίστηκαν 88 γονίδια τα οποία έχουν μεγάλη συνδεδιμότητα μεταξύ τους και σχηματίζουν ένα πυκνό δίκτυο πρωτεινικών αλληλεπιδράσεων (Εικόνα 4.6). Οι πρωτεΐνες αντιπροσωπεύονται ως κόμβοι και οι ακμές του δικτύου (γραμμές) είναι οι αλληλεπιδράσεις τους οι οποίες αντιστοιχούν σε διάφορους μοριακούς τρόπους δράσης τους. Οι πρωτεΐνες οι οποίες

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αλληλεπιδρούν με περισσότερους από έξι κόμβους (άλλες πρωτεινες) σε επίπεδο σημαντικότητας 0.7, μαζί με τις αντίστοιχες αλληλεπιδράσεις τους χρησιμοποιήθηκαν για περαιτέρω ανάλυση (Πίνακας 4.3). Αυτές οι πρωτεΐνες φαίνεται να σχηματίζουν δύο διακριτά υπο-δίκτυα όπως παρουσιάζεται στην Εικόνα 4.6. Το πρώτο υπο-δίκτυο περιλαμβάνει γονίδια που εμπλέκονται στη φλεγμονή, ενώ το δεύτερο περιέχει τις πρωτεΐνες που είναι υπεύθυνες για την επεξεργασία του RNA και των διαδικασιών μεταφοράς στον πυρήνα (nuclear import/export). Το πρώτο υπο-δίκτυο περιλαμβάνει τα γονίδια: ADORA3, ARRB2, CCL5, CXCL6, CXCR2 (IL8RB), CXCR7 (ACKR3), FPR2 (FPRL1) και GPER (GPR30), ενώ το δεύτερο περιέχει NUP37, NUP43, RAE1 και SRSF1 και τα σχετικά γονίδια CCAR1, CSTF3, SNRP40 ή SEH1L, SNJPN, MIOS και B9D2. Ένα τρίτο πολύ μικρότερο υπο-δίκτυο μπορεί επίσης να θεωρηθεί αυτό το οποίο αποτελείται μόνο από τα γονίδια NOTCH1, IGFR1 (MGC18216) και SPI1. Αξίζει να σημειωθεί ότι τα γονίδια SERPINA1, SERPINB1, WDR59, RBL1 και CTSG αποτελούν κεντρικούς κόμβους οι οποιοι συνδέουν τα πρώτα δύο υπο-δίκτυα. Η περαιτέρω ανάλυση των 88 γονίδιων έδειξε ότι μεταξύ αυτών υπάρχουν τρία σημαντικά βιοχημικά μοναπάτια KEGG (Πίνακας 4.4). Δύο από αυτά τα μονοπάτια έχουν σχέση με την ανοσία και τη φλεγμονή (ARRB2, CCL5, CXCL6, CXCR2 και CXCR7) και ένα μονοπάτι εμπλέκεται στη μεταφορά του RNA (NUP37, NUP43 και RAE1).

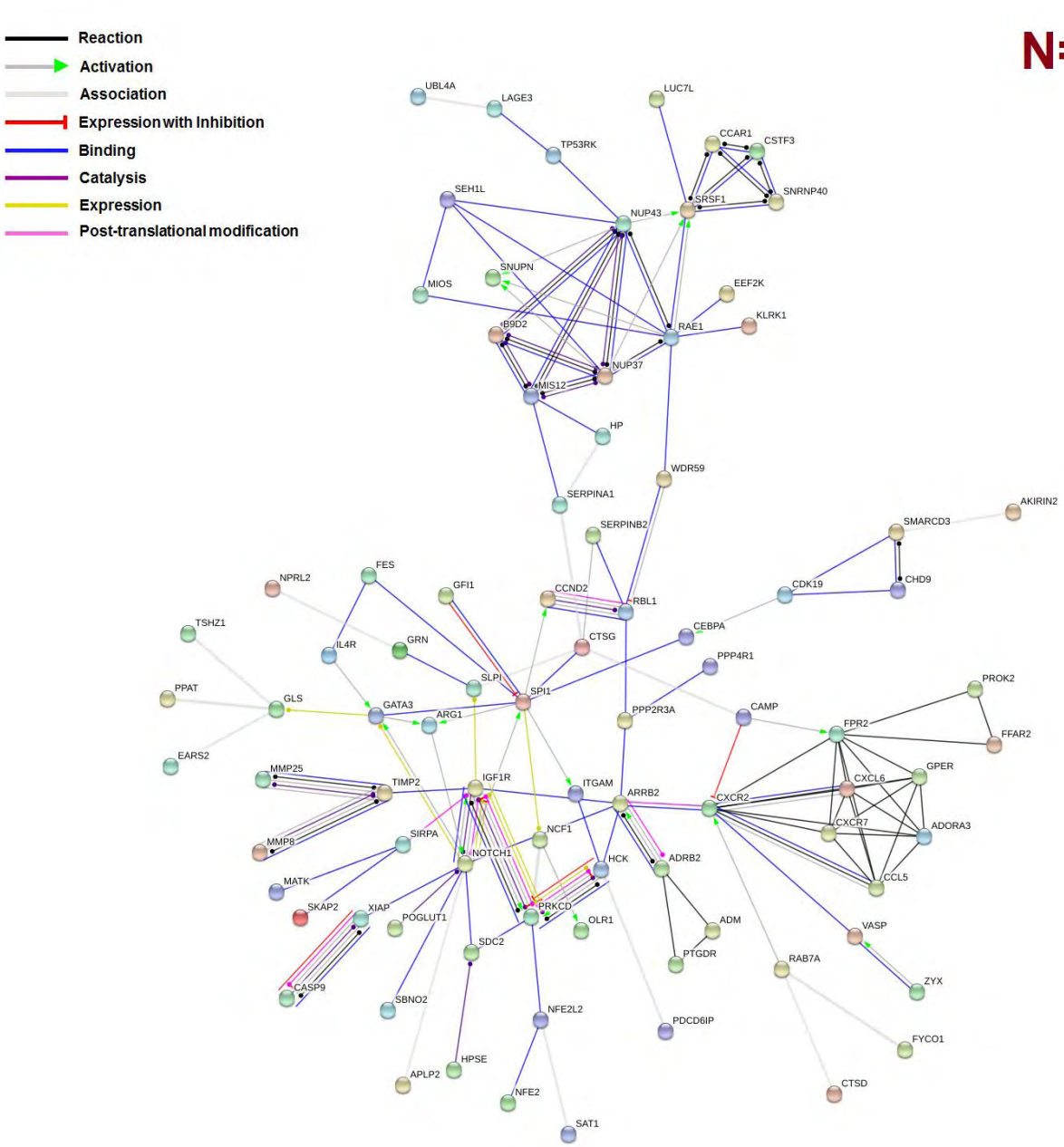
Επιπλέον, πραγματοποιήθηκε μια σύγκριση των 626 διαφορικά εκφρασμένων γονιδίων με στόχο να εντοπιστεί αν κάποιο από αυτά τα γονίδια έχει χρησιμοποιηθεί ως παράγοντας γενετικού κινδύνου από μελέτες γενετικής συσχέτισης (δηλαδή, με πολυμορφισμούς του να είναι γνωστό ότι επηρεάζουν την ασθένεια). Για τη σύγκριση χρησιμοποιήθηκε η λίστα των συσχετίσεων γονιδίων και ασθενειών που δημιουργήθηκε και παρουσιάζεται εκτενώς στο επόμενο Κεφάλαιο (Kontou et al., 2016d) και περιλαμβάνει πληροφορίες από τρεις διαφορετικές βάσεις γενετικών δεδομένων. Από την αναζήτηση προέκυψε ότι μόνο 8 γονίδια είναι κοινά: FES, GPD1L, IMPA2, OLR1, PGS1, PPP1R3B, PPP1R3B, ST3GAL4. Είναι ενδιαφέρον ότι τα γονίδια αυτά δεν συνδέονται μεταξύ τους σύμφωνα με το δίκτυο αλληλεπιδράσεων της STRING. Είναι σημαντικό, επίσης, να σημειωθεί ότι τρία από αυτά τα γονίδια (FES, ST3GAL4 και PPP1R3B) συμπεριλαμβάνονται στα κορυφαία (top) 60 γονίδια με FDR μικρότερο από 10^{-8} (Πίνακας 4.4).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 4.5. Το συνολικό δίκτυο αλληλεπιδράσεων πρωτεϊνών-πρωτεϊνών των διαφορικά εκφρασμένων γονίδιων στο έμφραγμα του μυοκαρδίου.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 4.6. Δίκτυο αλληλεπιδράσεων πρωτεϊνών-πρωτεϊνών που κωδικοποιούνται από 88 διαφορικά εκφρασμένα γονίδια, τα οποία εμφανίζουν ισχυρή συνδεσιμότητα, στο έμφραγμα του μυοκαρδίου. Οι γραμμές με τα διαφορετικά χρώματα δείχνουν τα διαφορετικά είδη των αλληλεπιδράσεων.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 4.3. Τα γονίδια με τη μεγαλύτερη συνδεσιμότητα στο δίκτυο αλληλεπιδράσεων πρωτεϊνών-πρωτεϊνών σε δυο διαφορετικά επίπεδα σημαντικότητας (0.7 και 0.9).

Γονίδιο	**Όνομα Γονιδίου**	**Συνδεδεμένοι κόμβοι (0.7)**	**Συνδεδεμένοι κόμβοι (0.9)**
CXCR2 (IL8RB)	C-X-C motif chemokine receptor 2	10	9
NOTCH1	notch 1	10	5
SPI1	Spi-1 proto-oncogene	10	5
FPR2 (FPRL1)	formyl peptide receptor 2	9	9
RAE1	ribonucleic acid export 1	9	4
NUP43	nucleoporin 43	8	6
NUP37	nucleoporin 37	7	7
SRSF1	serine and arginine rich splicing factor 1	7	6
CXCR7 *(ACKR3)*	chemokine (C-X-C motif) receptor 7 *(atypical chemokine receptor 3)*	6	6
CCL5	C-C motif chemokine ligand 5	6	6
CXCL6	C-X-C motif chemokine ligand 6	6	6
ADORA3	adenosine A3 receptor	6	6
GPER (GPR30)	G protein-coupled estrogen receptor 1	6	6
ARRB2	arrestin beta 2	6	5
IGF1R (MGC18216)	insulin like growth factor 1 receptor	6	3

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

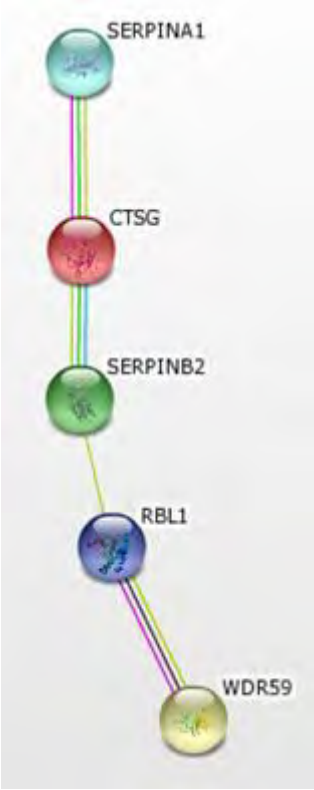
Πίνακας 4.4. Τα βιοχημικά μονοπάτια της KEGG στα οποία συμμετέχουν τα γονίδια με τη μεγαλύτερη συνδεσιμότητα, *adjP ≤ 0.001.

Μονοπάτι KEGG	**Όνομα Γονιδίου**	**adjP**
Cytokine-cytokine receptor interaction	CCL5 CXCR7 CXCR2 CXCL6	7.80e-05
Chemokine signaling pathway	CCL5 ARRB2 CXCR2 CXCL6	0.0003
RNA transport	NUP43 NUP37 RAE1	0.0003

Πίνακας 4.4. Τα top-60 γονίδια με FDR $<10^{-8}$.

MMP25	HCK	UST
S100A8	SLC16A3	C18ORF17
QIL1	OACT1	FLJ10081
ADRB2	LOC651738	ST3GAL4
CEACAM3	NCALD	TRIB2
CAMP	PLEKHG4	CCDC76
SLC19A1	PAFAH2	TAF15
LAX1	LOC653610	ALPL
DXS9879E	EIF3S7	TCN1
LOC649986	MGC7036	WDR57
CHST12	ZYX	CRYZL1
ATP6V0E2L	UBL4A	KIAA0701
PGLYRP1	LOC652878	PPP1R3B
CKAP1	HMFN0839	PTPNS1
FES	CTSD	KLRK1
HK3	LOC650761	TRAF3IP3
PDXP	MRPS5	LAT
RBL1	TRA16	IBRDC2
EVA1	ADM	SNAPC3
HMG20A	ANXA3	LOC440926

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 4.7. Το μονοπάτι αλληλεπίδρασης των 5 γονιδίων που αποτελούν κεντρικούς κόμβους και συνδέουν τα υπο-δίκτυα του συνολικού δικτύου των διαφορικά εκφρασμένων γονιδίων.

4.4.4 Συμπεράσματα

Στην παρούσα μελέτη, συνδυάζονται για πρώτη φορά όλα τα διαθέσιμα σύνολα δεδομένων μικροσυστοιχιών σχετικά με την εμφάνιση του εμφράγματος του μυοκαρδίου και πραγματοποιήθηκε μια μετα-ανάλυση προκειμένου να εντοπιστούν τα διαφορικά εκφρασμένα γονίδια που μπορούν να χρησιμοποιηθούν ως βιοδείκτες για την πρόγνωση εμφάνισης της ασθένειας. Ένα από τα κύρια προβλήματα που αφορούν τα πειράματα μικροσυστοιχιών είναι η έλλειψη τυποποίησης με αποτέλεσμα, τα δεδομένα που συλλέγονται από διαφορετικές πλατφόρμες μικροσυστοιχιών να μην μπορούν να συγκριθούν ή να αναπαραχθούν με ακρίβεια (Ioannidis et al., 2009). Αυτό οφείλεται κυρίως στην έλλειψη δεδομένων και στον ελλιπή σχολιασμό τους ή στην έλλειψη προδιαγραφών σχετικά με την επεξεργασία και την ανάλυση. Το θέμα της σύγκρισης των δεδομένων που παράγονται από

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

διαφορετικές πλατφόρμες είναι υπό διερεύνηση εδώ και αρκετό καιρό (Jarvinen et al., 2004) και το φιλτράρισμα των ανιχνευτών έχει αποδειχθεί ότι βελτιώνει σημαντικά τη σύγκριση των δεδομένων εντός της ίδιας πλατφόρμας (Hwang et al., 2004).

Οι μέθοδοι για το συνδυασμό διαφορετικών συνόλων δεδομένων σε μια μετα-ανάλυση μπορούν να βοηθήσουν τους ερευνητές σε ορισμένα από τα προβλήματα που αναφέρθηκαν παραπάνω (Moreau et al., 2003). Ωστόσο, τα θέματα που προαναφέρθηκαν, όπως η έλλειψη τυποποίησης, παραμένουν σημαντικά εμπόδια για την ανάπτυξη τέτοιων μεθόδων. Στην πρόσφατη βιβλιογραφία, υπάρχουν διάφορες μελέτες που συγκρίνουν τις διαφορετικές μεθόδους μετα-ανάλυσης (Campain and Yang, 2010; Chang et al., 2013; Hong and Breitling, 2008). Αξίζει να σημειωθεί ότι η έλλειψη τυποποίησης είναι επίσης εμφανής στην βιβλιογραφία σχετικά με μελέτες μετα-ανάλυσης των μικροσυστοιχιών, δεδομένου ότι διαφορετικές μέθοδοι και συνδυασμοί αυτών των μεθόδων έχουν χρησιμοποιηθεί στην πρόσφατη βιβλιογραφία. Μια πρόσφατη συστηματική έρευνα, είχε ως αποτέλεσμα την εμπειρική αξιολόγηση των άρθρων που αναφέρονται στη μετα-ανάλυση μικροσυστοιχιών (Tseng et al., 2012). Τα αποτελέσματα αυτής της αξιολόγησης ήταν πολύ ενδιαφέροντα, δεδομένου ότι, ένα μεγάλο μέρος από τις δημοσιευμένες μελέτες μετα-ανάλυσης διεξήχθη χρησιμοποιώντας την κατά κάποιον τρόπο ακατάλληλη μέθοδο μετα-ανάλυσης στην οποία συγκεντρώνονται μαζί τα αποτελέσματα των επιμέρους μελετών και εξάγεται από εκεί η τελική εκτίμηση χωρίς να συνυπολογίζεται το βάρος της κάθε μελέτης στο τελικό αποτέλεσμα (pooling method). Αυτό είναι ένα πολύ γνωστό πρόβλημα στη βιβλιογραφία σχετικά με τη μετα-ανάλυση και αυτή η προσέγγιση της συγκέντρωσης των συνόλων δεδομένων, προκειμένου να δημιουργηθεί απλά ένα μεγαλύτερο σύνολο δεν συνιστάται, δεδομένου ότι μπορεί να οδηγήσει σε διάφορους τύπους συστηματικού σφάλματος (bias).

Με την εφαρμογή αξιόπιστων στατιστικών μεθοδολογιών, εντοπίσαμε 626 στατιστικά σημαντικά γονίδια. Αξίζει να σημειωθεί ότι η μετα-ανάλυση αποκάλυψε περισσότερα γονίδια ως στατιστικά σημαντικά από την ένωση των αποτελεσμάτων των επιμέρους μελετών. Επειδή μια μετα-ανάλυση προσθέτει στατιστικό βάρος και ισχύ για την εκτίμηση του αποτελέσματος, το γεγονός ότι εντοπίστηκαν περισσότερα στατιστικά σημαντικά γονίδια σχετικά με την εμφάνιση του εμφράγματος του μυοκαρδίου ενισχύει την ευρωστία και την αξία της μετα-ανάλυσης στον τομέα της ανάλυσης δεδομένων υψηλής απόδοσης. Με βάση την βιοπληροφορική ανάλυση πετύχαμε την οπτικοποίηση αυτών των γονιδιακών αλληλεπιδράσεων, την ταυτοποίηση των βιοχημικών οδών και των βιολογικών διαδικασιών στις οποίες συμμετέχουν.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Είναι σημαντικό να σημειωθεί ότι τα γονίδια που βρέθηκαν στατιστικά σημαντικά για το έμφραγμα στην παρούσα μελέτη δεν συμπίπτουν, τουλάχιστον στο μεγαλύτερο μέρος τους, με γονίδια πολυμορφισμοί των οποίων έχουν προσδιοριστεί στο παρελθόν από τις κλασικές μελέτες γενετικής συσχέτισης ή τις ευρυγονιδιωματικές μελέτες. Μόνο οκτώ γονίδια είναι κοινά και σύμφωνα με την ανάλυση από τη STRING και από τις βάσεις γενετικών δεδομένων τα γονίδια αυτά εμπλέκονται στο μεταβολισμό των λιπιδίων (GPD1L, OLR1, PGS1), στο μεταβολισμό του γλυκογόνου και της γλυκόζης και έτσι συνδέονται με τον σακχαρώδη διαβήτη (PPP1R3B, ST3GAL4) ή εμπλέκονται σε διαδικασίες μεταφοράς στη μεμβράνη και σηματοδότησης (FES, IMPA2 και ABCB1). Είναι σημαντικό ότι τα τρία από αυτά τα γονίδια FES, ST3GAL4 και PPP1R3B είναι στα κορυφαία (top) 60 γονίδια με το μικρότερο p-value επομένως με την πιο ισχυρή συσχέτιση. Το γεγονός ότι η μέθοδός μας εντόπισε γονίδια που εμπλέκονται στο μεταβολισμό των λιπιδίων και του σακχάρου που είναι ευρέως γνωστοί ως γενετικοί παράγοντες κινδύνου για την εμφάνιση του εμφράγματος του μυοκαρδίου επιβεβαιώνει την αξιοπιστία της μετα-ανάλυσή που πραγματοποιήθηκε. Αν και τα κοινά γονίδια ήταν πολύ λίγα, η κατάσταση αυτή αντανακλά τα διαφορετικά επίπεδα πολυμορφισμών των γονιδίων και των πρωτεϊνών που εκφράζονται στην ανάπτυξη των καρδιαγγειακών βλαβών και εξελίσσονται σε έμφραγμα του μυοκαρδίου.

Με βάση τη βιοπληροφορική ανάλυση που διεξήχθη, μια κυρίαρχη ομάδα 88 στενά συνδεδεμένων γονιδίων ανακαλύφθηκε από τα διαφορικά εκφρασμένα γονίδια στο έμφραγμα. Η ανάλυση των μεταβολικών μονοπατιών τα κατατάσσει σε φλεγμονώδεις και θρομβωτικές διαδικασίες, σε διαδικασίες επούλωσης των πληγών και διαδικασίες μεταφοράς του RNA. Επιπλέον, αυτό επιβεβαιώθηκε και από το εργαλείο biocompedium στο οποίο είδαμε τα γονίδια να εμπλέκονται σε διαδικασίες απόκρισης στο στρες και στη φλεγμονή. Σε αυτό το σημείο καθίσταται σημαντικό να αναφερθεί ότι το έμφραγμα του μυοκαρδίου είναι κυρίως αποτέλεσμα της αθηροσκλήρωσης, μιας κατάστασης που εκδηλώνεται με χρόνια φλεγμονώδη απόκριση των λευκών αιμοσφαιρίων στα τοιχώματα των αρτηριών (Bonaventura et al., 2016).

Τέλος, θα πρέπει να αναφερθούμε στα πέντε γονίδια που αποτελούν κεντρικούς κόμβους και συνδέουν τα δύο μεγάλα υπο-δίκτυα, το δίκτυο των φλεγμονωδών γονιδίων υποδοχέων- κυτοκίνης και το δίκτυο των γονίδιων μεταφοράς. Τα γονίδια αυτά είναι: SERPINA1, SERPINB2, WDR59, RBL1 και CTSG και είναι γραμμικά συνδεμένα μεταξύ τους στο ίδιο «μονοπάτι». Δύο από αυτά τα γονίδια είναι αναστολείς πεπτιδάσης της σερπίνης (SERPINA1, SERPINB2), ενώ το CTSG είναι μια πρωτεάση σερίνης με ειδικότητα τύπου θρυψίνης και χυμοθρυψίνης. Το γονίδιο WDR59 είναι ένα συστατικό του συμπλέγματος GATOR το οποίο λειτουργεί ως ενεργοποιητής της βιοχημικής οδού TORC1. Το γονίδιο

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

RBL1, εμπλέκεται στη ρύθμιση της έναρξης της κυτταρικής διαίρεσης. Αξιοσημείωτο είναι ότι ο πολυμορφισμός 4G/5G του γονιδίου της SERPINE1, ενός άλλου αναστολέα πεπτιδάσης της σερπίνης, έχει βρεθεί να σχετίζεται σημαντικά με έμφραγμα του μυοκαρδίου και με την αγγειακή θρόμβωση (Tsantes et al., 2007). Το γεγονός ότι η μέθοδός μας ανίχνευσε μερικά γονίδια τα οποία είναι ήδη γνωστά ότι σχετίζονται με το έμφραγμα του μυοκαρδίου ενισχύει τα νέα αποτελέσματα αυτής της μελέτης η οποία είχε ως στόχο την εύρεση γονιδίων που εμπλέκονται στο έμφραγμα του μυοκαρδίου.

4.5 Δημοσιεύσεις σε επιστημονικά περιοδικά

Στο Κεφάλαιο αυτό παρουσιάστηκε μια ανασκόπηση όλων των μεθοδολογιών που χρησιμοποιούνται για τη μετα-ανάλυση δεδομένων μικροσυστοιχιών με σκοπό την εύρεση διαφορικά εκφρασμένων γονιδίων από περισσότερες από μια μελέτες που ερευνούν το ίδιο βιολογικό ερώτημα, ενώ προτάθηκαν και σημαντικές τροποποιήσεις που παρουσιάζουν μια σειρά από πλεονεκτήματα. Η ανασκόπηση των μεθόδων μετα-ανάλυσης και η υλοποίηση των μεθόδων μετα-ανάλυσης που παρουσιάστηκαν στο παρόν Κεφάλαιο έχει γίνει δεκτή ως άρθρο ανασκόπησης στο *Methods in Molecular Biology* (Kontou et al., 2016a). Επιπλέον, η μετα-ανάλυση των δεδομένων μικροσυστοιχιών με σκοπό την εύρεση γονιδίων που εκφράζονται διαφορικά στην εμφάνιση του εμφράγματος του μυοκαρδίου η οποία παρουσιάστηκε στην ενότητα 4.4 έχει υποβληθεί ήδη σε διεθνές επιστημονικό περιοδικό (Kontou et al., 2016b).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 5

Ανάλυση δικτύου γονιδίων και των συσχετίσεών τους με ασθένειες

5.1 Εισαγωγή

Οι ανθρώπινες γενετικές διαταραχές σπάνια αποδίδονται στη δράση ενός μόνο γονιδίου, συνήθως αποδίδονται στη συνδυαστική δράση περισσότερων του ενός (Cordell and Clayton, 2005; Goldstein, 2009). Κατά την τελευταία δεκαετία, μέθοδοι αλληλούχισης μεγάλης-κλίμακας (high-throughput sequencing) όπως μελέτες σάρωσης γονιδιώματος (genome-wide association studies ή GWAS) επιτρέπουν την ανίχνευση μεγάλου αριθμού γονιδίων που εμπλέκονται σε ασθένειες και εκατομμυρίων μονονουκλεοτιδικών πολυμορφισμών (SNPs) που σχετίζονται με ασθένειες (Hirschhorn, 2009; Manolio, 2010).

Δίκτυα που απεικονίζουν συσχετίσεις ασθενειών-γονιδίων επιτρέπουν όχι μόνο τη διερεύνηση της γενετικής πολυπλοκότητας μιας συγκεκριμένης ασθένειας, αλλά και τη σχέση μεταξύ φαινομενικά διαφορετικών ασθενειών (Barabasi et al., 2011; Pawson and Linding, 2008). Επιπλέον, δίκτυα αλληλεπιδράσεων γονιδίων και ασθενειών επιτρέπουν την πρόβλεψη γονιδίων σχετιζόμενων με ασθένειες, την αποσαφήνιση των μηχανισμών που εμπλέκονται στην παθογένεια των ασθενειών και το σχεδιασμό θεραπευτικών στρατηγικών (Barabasi et al., 2011; Pawson and Linding, 2008). Έχει αποδειχθεί ότι τα γονίδια που σχετίζονται με ίδιες ή παρόμοιες ασθένειες έχουν μεγαλύτερη τάση να αλληλεπιδρούν μεταξύ τους (Hartwell et al., 1999; Oti and Brunner, 2007). Σε αντίθεση με τις ασθένειες με διαφορετικούς φαινοτύπους, έχει παρατηρηθεί ότι οι ασθένειες με παρόμοιους φαινοτύπους έχουν κοινά γονίδια (Goh et al., 2007).

Στην παρούσα μελέτη, κατασκευάστηκε ένα δίκτυο αλληλεπιδράσεων ανθρώπινων γονιδίων και ασθενειών, προκειμένου να διερευνηθούν οι συσχετίσεις μεταξύ των γενετικών ασθενειών του ανθρώπου και των γονιδίων που σχετίζονται με τις ασθένειες αυτές. Ο Goh και οι συνεργάτες του κατασκεύασαν το 2007 (Goh et al., 2007) ένα παρόμοιο δίκτυο που βασίζεται σε συσχετίσεις ασθενειών-γονιδίων οι οποίες έχουν εξαχθεί από τη βάση δεδομένων OMIM (Online Mendelian Inheritance in Man) (Amberger et al., 2015). Ωστόσο, αν και η OMIM αποτελεί την κύρια βάση δεδομένων γενετικών δεδομένων Μεντελικών ασθενειών, περιλαμβάνει μόνο τις πιο σημαντικές και σπάνιες ασθένειες καθιστώντας τη μελέτη που διεξήχθη από τον Goh και τους συνεργάτες του μάλλον ελλιπή. Προκειμένου να καλυφθεί αυτό το κενό και να συμπεριληφθούν περισσότερες ασθένειες

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

πολυπαραγοντικής αιτιολογίας, η μελέτη μας επεκτάθηκε, περιλαμβάνοντας δεδομένα από δύο διαφορετικές βάσεις δεδομένων, την GAD (Becker et al., 2004) και τον κατάλογο NHRI GWAS (Welter et al., 2014). Η GAD περιλαμβάνει δεδομένα γενετικής συσχέτισης για διάφορες πολυπαραγοντικές ασθένειες (κυρίως ασθένειες με μέτρια ως μικρή διεισδυτικότητα), αντλώντας την πληροφορία απευθείας από την PUBMED. Ο κατάλογος NHGRI GWAS περιλαμβάνει μια συλλογή δημοσιευμένων GWAS για τις συσχετίσεις SNP-ασθενειών. Στη βάση αυτή περιέχονται κυρίως πολυπαραγοντικές ασθένειες μεγάλης επίπτωσης, ενώ σε αντίθεση με την GAD, οι γενετικοί πολυμορφισμοί που μελετώνται, δεν είναι επιλεγμένοι εκ των προτέρων αλλά προκύπτουν από ανάλυση όλου του γονιδιώματος. Ως εκ τούτου, οι βάσεις αυτές, περιέχουν σημαντικές πληροφορίες που δεν υπάρχουν στην OMIM και θα πρέπει να ληφθούν υπόψη.

5.2 Μέθοδοι

5.2.1 Συλλογή Δεδομένων

Τα δεδομένα των συσχετίσεων γονιδίων-ασθενειών συλλέχθηκαν από τρεις βάσεις δεδομένων οι οποίες είναι διαθέσιμες στο ευρύ κοινό. Αυτές είναι:

i) Η βάση δεδομένων OMIM (Online Mendelian Inheritance in Man) (Amberger et al., 2015) που περιέχει πληροφορίες για τις γενετικά κληρονομήσιμες ασθένειες. Τα δεδομένα της βάσης δεδομένων OMIM περιέχονται στο αρχείο «genemap2.txt» το οποίο ανακτήθηκε από τη διεύθυνση https://omim.org/ στις 17 Αυγούστου 2013. Επιλέχθηκαν οι συσχετίσεις ασθενειών-γονιδίων, τύπου 3 (το οποίο σημαίνει ότι είναι γνωστή η μοριακή βάση της ασθένειας), και κατόπιν δημιουργήθηκε ένα αρχείο από αυτά τα δεδομένα, στο οποίο μία στήλη περιέχει τα ονόματα των ασθενειών και μια άλλη στήλη τα ονόματα των γονιδίων που σχετίζονται με τις ασθένειες. Στην περίπτωση που το όνομα του γονιδίου δεν ήταν διαθέσιμο, η αντίστοιχη γενωμική περιοχή στην οποία χαρτογραφείται το γονίδιο περιελήφθη στη μελέτη αντ' αυτού.

ii) Η βάση δεδομένων GAD (Becker et al., 2004) που περιέχει πληροφορίες για πολυπαραγοντικές ασθένειες. Τα δεδομένα της GAD περιέχονται στο αρχείο "all.txt" και ανακτήθηκαν από τη διεύθυνση http://geneticassociationdb.nih.gov στις 17 Αυγούστου 2013. Επιλέχθηκαν οι συσχετίσεις ασθενειών-γονιδίων, υπό τον τίτλο «Association» δηλαδή την ύπαρξη συσχέτισης. Λόγω του τεράστιου όγκου των διαθέσιμων συσχετίσεων στην GAD και του γεγονότος ότι πολλές από τις δημοσιευμένες μελέτες περιέχουν και «αρνητικά»

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αποτελέσματα, επιλέχθηκαν μόνο οι συσχετίσεις για τις οποίες η GAD είχε την ένδειξη «significant». Επιπλέον, επειδή είναι γνωστό ότι οι μελέτες γενετικής συσχέτισης δίνουν πολλές φορές και αντικρουόμενα αποτελέσματα (Ioannidis et al., 2001), επιλέχθηκαν τελικά μόνο οι στατιστικώς σημαντικές συσχετίσεις που προέκυψαν από δημοσιευμένες μετα-αναλύσεις. Από αυτά τα δεδομένα, δημιουργήθηκε ένα αρχείο με τρεις στήλες που περιέχουν το όνομα της ασθένειας, το όνομα του γονιδίου και το γενετικό πολυμορφισμό που σχετίζεται με τη συγκεκριμένη ασθένεια. Όπως και πριν, σε περίπτωση που το όνομα του γονιδίου δεν ήταν διαθέσιμο, συμπεριελήφθη η αντίστοιχη χρωμοσωμική θέση του γονιδίου.

iii) Η βάση δεδομένων NHRI GWAS (Welter et al., 2014) που περιέχει συσχετίσεις πολυμορφισμών (SNPs) με ασθένειες. Τα δεδομένα της ανακτήθηκαν από τη διεύθυνση http://www.genome.gov/gwastudies στις 23 Αυγούστου 2013. Στη μελέτη συμπεριλήφθηκαν μόνο οι γονιδιακοί πολυμορφισμοί. Στη συνέχεια, δημιουργήθηκε ένα αρχείο που περιλαμβάνει το όνομα της ασθένειας, το όνομα του γονιδίου και το σχετιζόμενο γονιδιακό πολυμορφισμό (SNP).

Τα τρία αρχεία που προέκυψαν από τις βάσεις δεδομένων ενώθηκαν και προέκυψε ένα τέταρτο αρχείο το οποίο περιέχει όλες τις πληροφορίες σχετικά με τις συσχετίσεις γονιδίων και ασθενειών. Στο εξής ο όρος JOINT ή συνολικό δίκτυο αφορά το δίκτυο που έχει προκύψει από τη συνένωση των τριών βάσεων δεδομένων.

5.2.2 Ονοματολογία ασθενειών και γονιδίων

Ένα πρόβλημα που προέκυψε, ήταν η ετερογένεια στην ονοματολογία των ασθενειών στις τρεις βάσεις δεδομένων, η οποία έκανε δύσκολη τη σύγκριση μεταξύ των δεδομένων. Προκειμένου να διατηρηθεί μια ομοιογενής ονοματολογία και κατηγοριοποίηση των ασθενειών, χρησιμοποιήθηκε η ονοματολογία που περιγράφεται στην Βάση Δεδομένων International Classification of Diseases (ICD). Η Βάση Δεδομένων ICD (http://www.who.int/classifications/icd/en/), του Παγκόσμιου Οργανισμού Υγείας (WHO), αποτελεί ένα σύστημα ταξινόμησης των ασθενειών όπου παρόμοιες ασθένειες ομαδοποιούνται σε μεγάλες κατηγορίες. Το όνομα κάθε ασθένειας - για κάθε ένα από τα τρία σύνολα δεδομένων- χρησιμοποιήθηκε ως ερώτημα έναντι της ICD για την εύρεση σχετικού συνωνύμου. Στην περίπτωση που ένα κατάλληλο συνώνυμο για ένα συγκεκριμένο όνομα ασθένειας δεν βρέθηκε στην ICD, η ευρύτερη κατηγορία ασθενειών στην οποία ανήκει η ασθένεια επιλέχθηκε αντ' αυτού. Σε περίπτωση

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

που δεν βρέθηκε ούτε το συνώνυμο ούτε και η ευρύτερη κατηγορία μιας ασθένειας στην ICD, η εν λόγω ασθένεια δεν έχει συμπεριληφθεί στην παρούσα μελέτη. Όσον αφορά τα ονόματα των γονιδίων, χρησιμοποιήθηκαν τα σύμβολα γονιδίων όπως περιγράφονται στην HGNC (HUGO Gene Nomenclature Committee) (Gray et al., 2015), ούτως ώστε να διασφαλισθεί μια ομοιόμορφη ονοματολογία μεταξύ των τριών συνόλων δεδομένων.

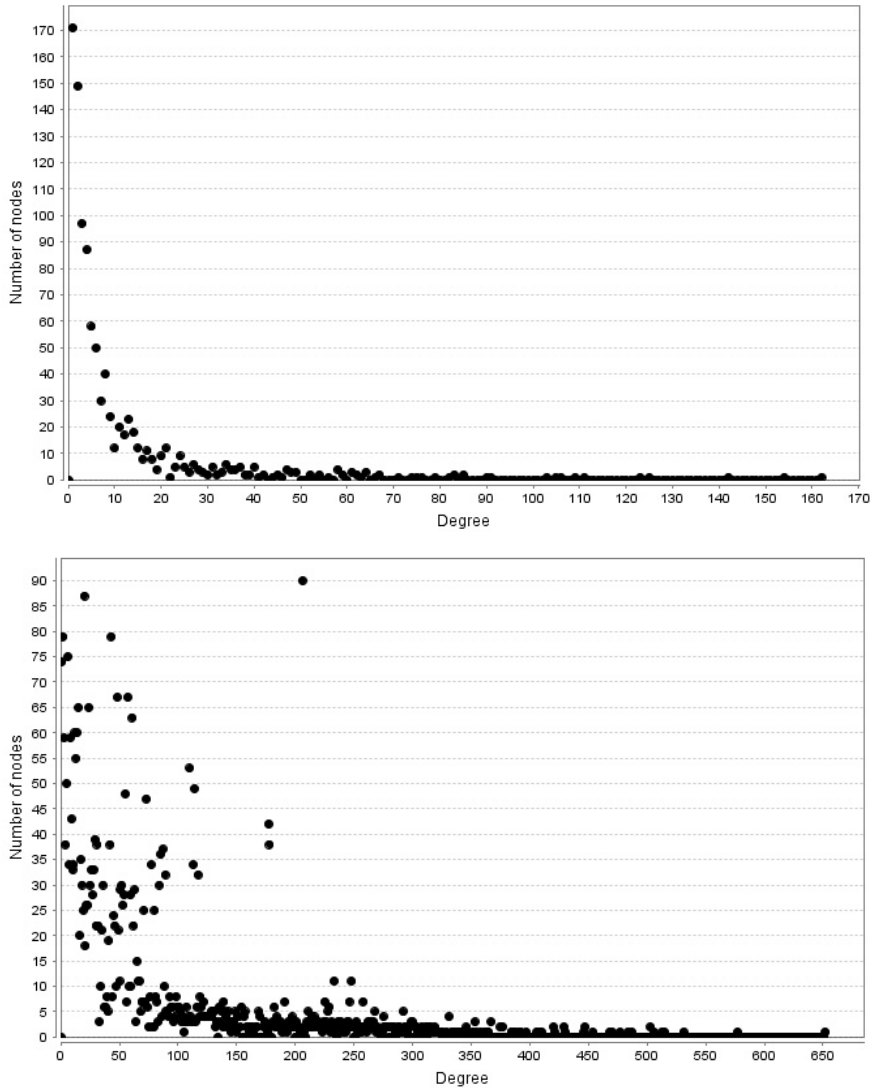
5.2.3 Ανάλυση Δικτύων και Οπτικοποίηση

Το λογισμικό Cytoscape v.3.2.1 (http://www.cytoscape.org/) χρησιμοποιήθηκε για τη στατιστική ανάλυση, επεξεργασία και οπτικοποίηση των δεδομένων του δικτύου. Οι συσχετίσεις ασθενειών-γονιδίων παρουσιάζονται ως διμερή δίκτυα, στα οποία δύο ομάδες κόμβων, που αντιστοιχούν σε ασθένειες και γονίδια, συνδέονται με ακμές. Δύο μονομερή δίκτυα δημιουργήθηκαν από το διμερές δίκτυο, χρησιμοποιώντας ένα πρόγραμμα γραμμένο σε γλώσσα προγραμματισμού Perl. Στα μονομερή δίκτυα, δύο κόμβοι συνδέονται μόνο αν και οι δύο συνδέονται με τον ίδιο κόμβο στο αρχικό διμερές δίκτυο. Για τις ομαδοποιήσεις στα μονομερή δίκτυα, χρησιμοποιήθηκε ο αλγόριθμος MCL (Markov Clustering) (Enright et al., 2002) που εφαρμόζεται στο Cytoscape (Cline et al., 2007).

Στα συνολικά μονομερή (monopartite) δίκτυα η ανάλυση επικεντρώνεται κυρίως στα γονίδια/ασθένειες που είναι περισσότερο συνδεδεμένα/συνδεδεμένες (έχουν το μεγαλύτερο βαθμό), δηλαδή αλληλεπιδρούν, με μεγάλο αριθμό γονιδίων/ασθενειών αντίστοιχα. Για το σκοπό αυτό, τέθηκε αυθαίρετα ένα όριο για τα top-20 πιο πολύ συνδεδεμένα γονίδια/ασθένειες που αντιστοιχούν στο άκρο της κατανομής του βαθμού των κόμβων (Εικόνα 5.1). Επιπλέον, μια σειρά διαδικασιών τυχαιοποίησης (randomization procedures) πραγματοποιήθηκε με σκοπό να διερευνηθεί η ευρωστία (robustness) των ευρημάτων. Έτσι, εξετάστηκε αν τα top-20 γονίδια/ασθένειες στο αρχικό συνολικό δίκτυο αντιπροσωπεύουν επίσης τους 20 πιο συνδεδεμένους κόμβος στα τυχαία δίκτυα. Αρχικά, στο συνολικό (JOINT) δίκτυο γονιδίων/ασθενειών έγινε τυχαία αντιμετάθεση 100 φορές στη στήλη με τις ασθένειες. Δημιουργήθηκαν 100 τυχαία διμερή δίκτυα και στη συνέχεια προέκυψαν 100 μονομερή δίκτυα γονίδιων-γονίδιων. Έπειτα, τυχαιοποιήθηκε και η στήλη των γονιδίων στο αρχικό συνολικό διμερές δίκτυο και δημιουργήθηκαν 100 τυχαία διμερή δίκτυα και στη συνέχεια 100 τυχαία δίκτυα ασθενειών - ασθενειών παρήχθησαν από τα τυχαία διμερή δίκτυα. Επιπλέον, χρησιμοποιήσαμε μια παρόμοια διαδικασία τυχαιοποίησης με ταυτόχρονη μετάθεση τόσο των ασθενειών όσο και των γονίδιων. Με τον τρόπο αυτό παρήχθησαν 100 τυχαία δίκτυα γονίδιων-γονίδιων και 100 δίκτυα ασθενειών - ασθενειών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Τέλος, προκειμένου να δοκιμαστεί η ευρωστία της τοπολογίας των τυχαιοποιημένων διμερών και μονομερών δικτύων, χρησιμοποιήθηκε ένας αλγόριθμος που πραγματοποιεί επανακαλωδίωση (rewiring) (Gobbi et al., 2014). Ο αλγόριθμος αυτός, διατηρώντας σταθερό το βαθμό των κόμβων του διμερούς δικτύου, επιτρέπει τον έλεγχο της κατανομής του βαθμού των κόμβων στα τυχαιοποιημένα δίκτυα γονιδίων-γονιδίων/ασθενειών–ασθενειών έτσι ώστε να είναι και εκεί σταθερός. Σε όλες τις διαδικασίες τυχαιοποίησης, οι κόμβοι κατατάχθηκαν ανάλογα με το βαθμό και ελέγχθηκε κατά πόσο τα top-20 γονίδια ή ασθένειες των 100 τυχαίων δικτύων ήταν όμοια με τα top-20 γονίδια ή ασθένειες του πραγματικού δικτύου.

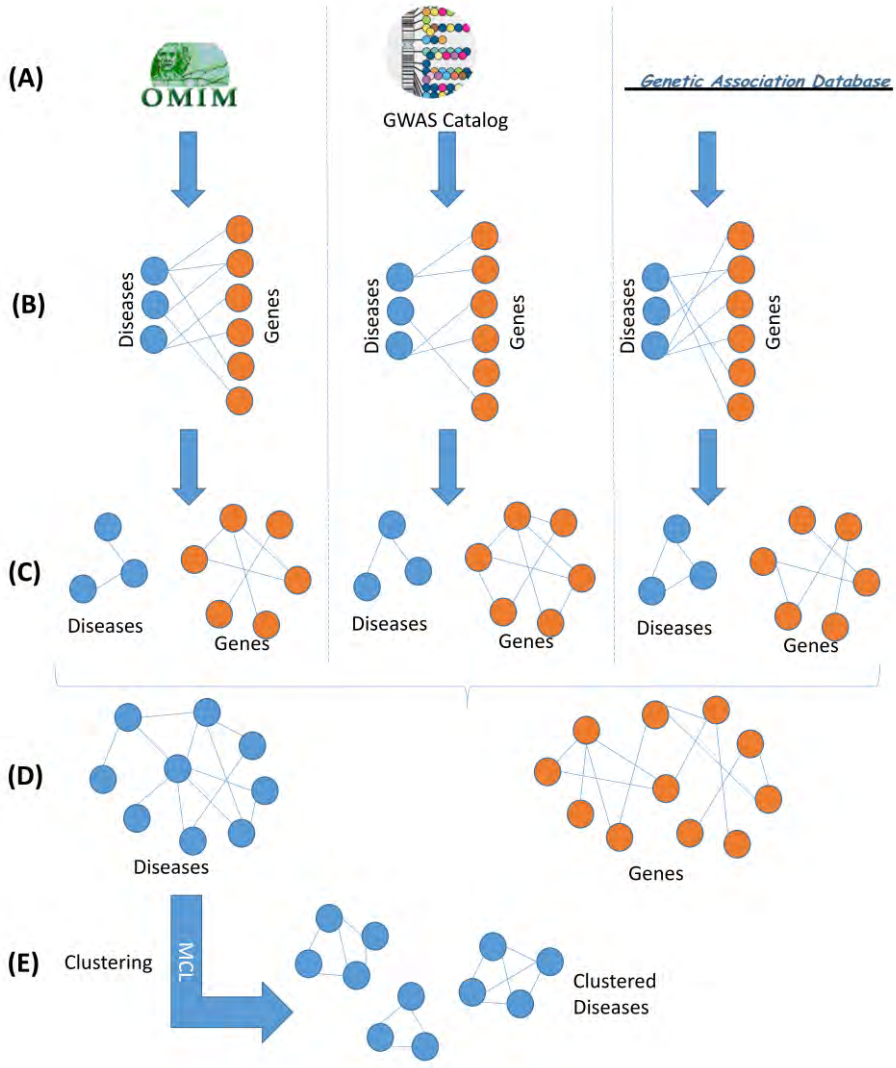


Εικόνα 5.1. Η κατανομή του βαθμού των κόμβων των γονιδίων (επάνω) και των ασθενειών (κάτω) στα συνολικά δίκτυα γονιδίων – γονιδίων και ασθενειών – ασθενειών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

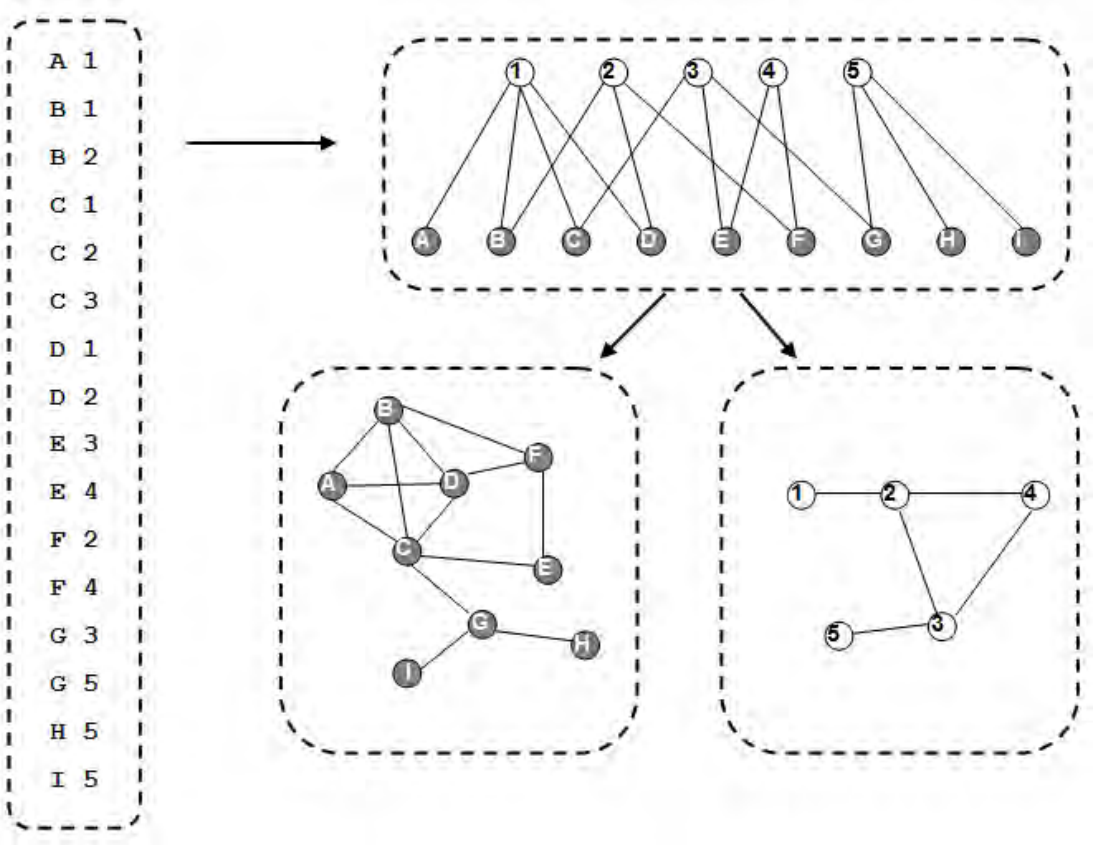
5.3 Αποτελέσματα

Στην παρούσα μελέτη, κατασκευάστηκαν μονομερή δίκτυα γονιδίων-γονιδίων και ασθενειών–ασθενειών από διμερή δίκτυα γονιδίων–ασθενειών μέσω της ενσωμάτωσης τριών βάσεων δεδομένων. Η διαδικασία συλλογής των δεδομένων αλλά και της ανάλυσης που ακολουθήθηκε στη μελέτη απεικονίζεται διαγραμματικά στην Εικόνα 5.2. Στην Εικόνα 5.3 απεικονίζεται η διαδικασία υλοποίησης του αλγόριθμου που μετατρέπει το διμερή γράφο με τις συσχετίσεις ασθενειών-γονιδίων σε μονομερή δίκτυα αλληλεπιδράσεων ασθενειών-ασθενειών και γονιδίων-γονιδίων. Ο αλγόριθμος αυτός υλοποιήθηκε σε γλώσσα προγραμματισμού Perl και είναι πλέον διαθέσιμος στο ευρύ κοινό στην ιστοσελίδα http://www.compgen.org/tools/powerclust.



Εικόνα 5.2. Διαγραμματική απεικόνιση της συλλογής δεδομένων και της ανάλυσης.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 5.3. Δημιουργία του δικτύου αλληλεπιδράσεων ασθενειών-ασθενειών και γονιδίων-γονιδίων, από το διμερές γράφο με τις συσχετίσεις ασθενειών-γονιδίων.

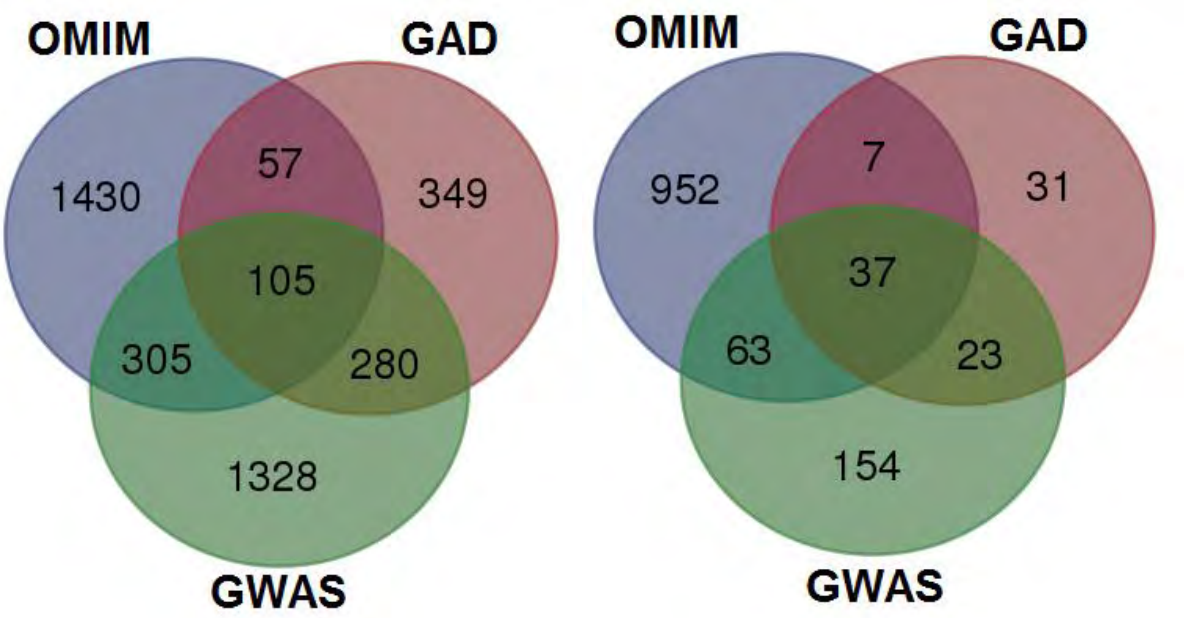
Τα μοναδικά ονόματα ασθενειών και γονιδίων, καθώς και οι μοναδικές συσχετίσεις ασθενειών-γονιδίων για κάθε μια από τις τρείς βάσεις δεδομένων παρουσιάζονται στον παρακάτω πίνακα:

	OMIM	**GAD**	**GWAS**	**Συνολικά**
Ονόματα Ασθενειών	1,059	98	277	1,267
Ονόματα Γονιδίων	1,897	791	2,018	3,854
Συσχετίσεις Ασθενειών-Γονιδίων	2,911	1,024	3,633	7,158

Τα γονίδια και οι ασθένειες από τις τρεις διαφορετικές βάσεις δεδομένων διαφέρουν σημαντικά μεταξύ τους, όπως απεικονίζεται στα παρακάτω διαγράμματα Venn. Αξίζει να σημειωθεί ότι και οι τρεις βάσεις δεδομένων έχουν μεταξύ τους μόνο 105 κοινά γονίδια και 37 κοινές ασθένειες. Το γεγονός

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

αυτό, δείχνει ότι η προσπάθεια συνένωσης δεδομένων από διαφορετικές πηγές, ήταν σωστή, καθώς σε διαφορετική περίπτωση, πολλές πιθανές συσχετίσεις θα είχαν αγνοηθεί. Πρέπει να τονιστεί εδώ, ότι η μέθοδος για την ενοποίηση, είχε διαφορετικό σκοπό από άλλες αντίστοιχες εργασίες, καθώς στόχος ήταν να συγκεντρωθούν όσο περισσότερα δεδομένα ήταν δυνατό (ενώ παραδοσιακά οι μεθοδολογίες μετα-ανάλυσης εστιάζουν σε μικρότερο ή μεγαλύτερο βαθμό, τη συμφωνία των διαφορετικών πηγών).



Εικόνα 5.4. Διάγραμμα Venn για την επικάλυψη των γονιδίων στις τρεις βάσεις δεδομένων (Αριστερά) και διάγραμμα Venn για την επικάλυψη των ασθενειών στις τρεις βάσεις δεδομένων (Δεξιά). Παρατηρούμε, ότι τόσο για τις ασθένειες, όσο και για τα γονίδια, η επικάλυψη είναι ιδιαίτερα μικρή.

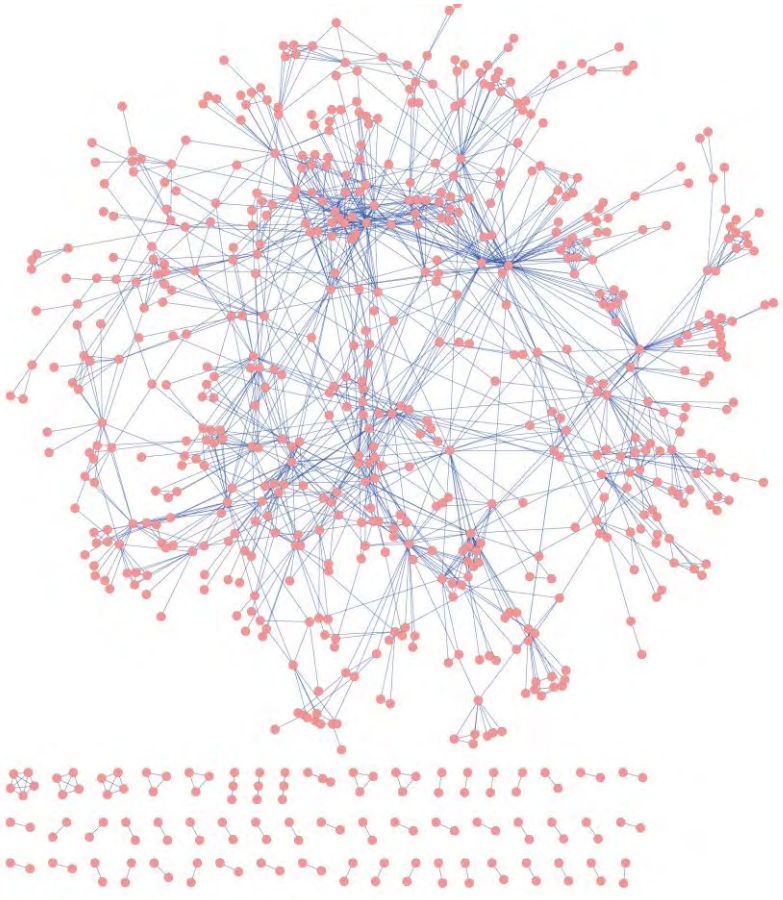
Αρχικά, τρεις μη κατευθυνόμενοι γράφοι κατασκευάστηκαν, ένας για κάθε σύνολο δεδομένων. Το ένα σύνολο κόμβων στους διμερείς γράφους αντιστοιχεί στις γενετικές ασθένειες και το άλλο σύνολο αντιστοιχεί στα γονίδια που σχετίζονται με τις ασθένειες αυτές. Στη συνέχεια κατασκευάστηκαν δύο σύνολα μη κατευθυνόμενων μονομερών γράφων. Το πρώτο σύνολο αποτελείται από τρία δίκτυα ασθενειών, ένα για κάθε βάση δεδομένων, όπου οι κόμβοι αντιπροσωπεύουν ασθένειες οι οποίες συνδέονται μεταξύ τους, εάν έχουν τουλάχιστον ένα κοινό γονίδιο που εμπλέκεται και στις δύο ασθένειες. Το δεύτερο σύνολο γράφων περιλαμβάνει τρία δίκτυα γονιδίων που σχετίζονται με ασθένειες. Οι κόμβοι του δικτύου αυτού αντιπροσωπεύουν γονίδια τα οποία συνδέονται μεταξύ τους μόνο αν συμμετέχουν στην ίδια ασθένεια. Επιπλέον, κατασκευάστηκαν δυο συνολικά δίκτυα – ασθενειών και γονιδίων- με συνένωση των τριών δικτύων ασθενειών και γονιδίων, αντίστοιχα.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πραγματοποιήθηκε ανάλυση όλων των δικτύων (8) με το πρόγραμμα ανάλυσης δικτύων Cytoscape. Παρακάτω παρατίθενται τα δίκτυα που προέκυψαν από την ανάλυση. Τόσο για το δίκτυο γονιδίων-γονιδίων όσο και για το δίκτυο ασθενειών-ασθενειών, παρατίθεται το δίκτυο της OMIM, της GAD, της GWAS, αλλά και το συνολικό το οποίο προκύπτει από τη συνένωση των τριών.

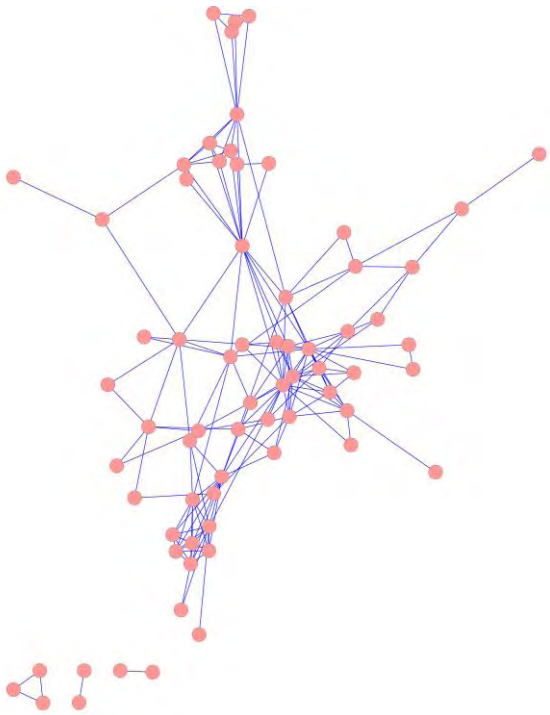
5.3.1 Δίκτυα ασθενειών -ασθενειών

Το πρώτο σύνολο μονομερών δικτύων αποτελείται από τέσσερα δίκτυα ασθενειών, της OMIM, GAD, GWAS και το συνολικό JOINT δίκτυο (Εικόνες 5.5-5.8). Οι κόμβοι του δικτύου είναι οι ασθένειες και οι ακμές του (αλληλεπιδράσεις) είναι η συσχέτισή τους με κάποιο γονίδιο.

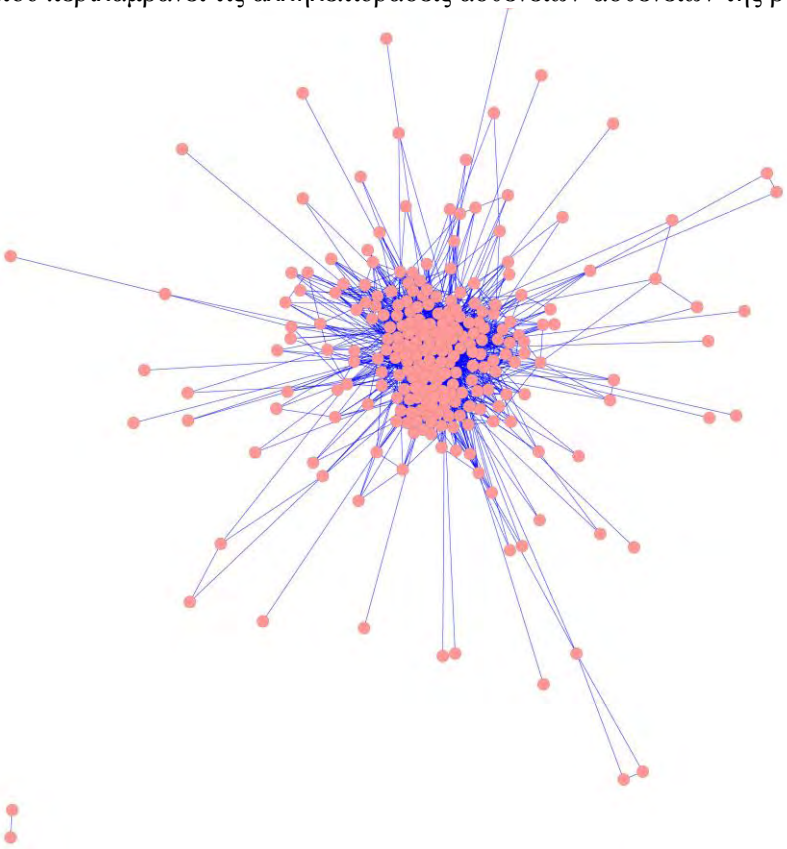


Εικόνα 5.5. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις ασθενειών-ασθενειών της βάσης δεδομένων OMIM.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

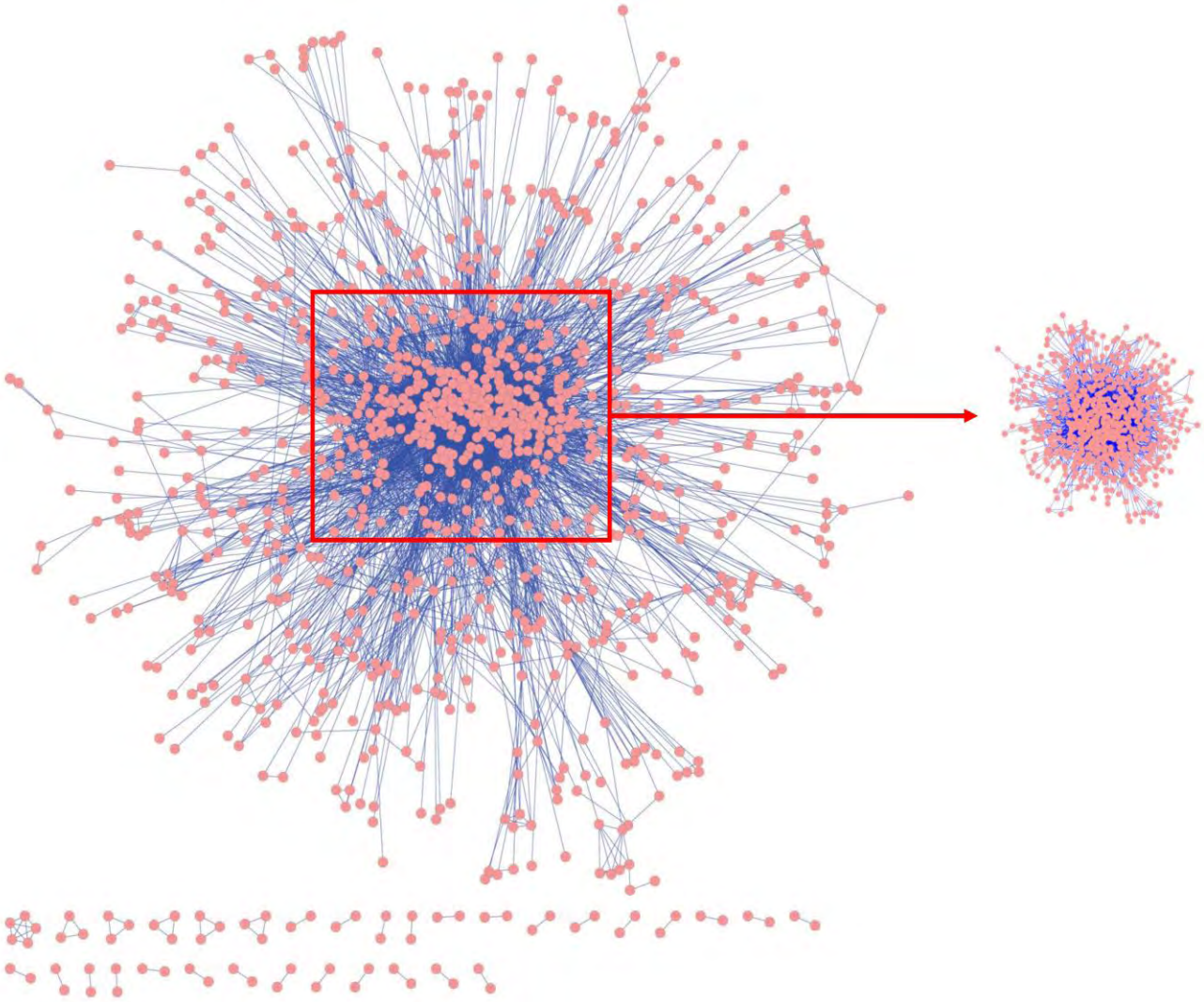


Εικόνα 5.6. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις ασθενειών-ασθενειών της βάσης δεδομένων GAD.



Εικόνα 5.7. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων της βάσης δεδομένων GWAS.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 5.8. Το συνολικό δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων.

Από τη στατιστική ανάλυση των δικτύων προέκυψε ότι τα δίκτυα ασθενειών – ασθενειών ανήκουν στην κατηγορία των δικτύων άνευ κλίμακας (scale-free networks) στα οποία η πιθανότητα ένας κόμβος να αλληλεπιδρά με ένα σύνολο k κόμβων ακολουθεί την κατανομή νόμου δύναμης (power law distribution). Η κατανομή του βαθμού των κόμβων $P(k)$ του δικτύου που έχει k συνδέσεις με άλλους κόμβους ορίζεται από τη σχέση $P(k) \sim k^{-\gamma}$, όπου γ η εκτιμώμενη παράμετρος του νόμου δύναμης για την κατανομή του βαθμού των κόμβων (Barabasi, 2009; Barabasi and Albert, 1999). Στα δίκτυα ασθενειών υπάρχουν λίγοι κόμβοι οι οποίοι εμφανίζουν μεγάλο αριθμό συνδέσεων (κεντρικοί κόμβοι). Επίσης, όσο πιο μικρή είναι η τιμή του εκθέτη του βαθμού (γ) τόσο περισσότερο οι ιδιότητες του δικτύου καθορίζονται από τους κεντρικούς κόμβους. Οι τιμές του γ στον Πίνακα 5.1 που είναι αρκετά

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

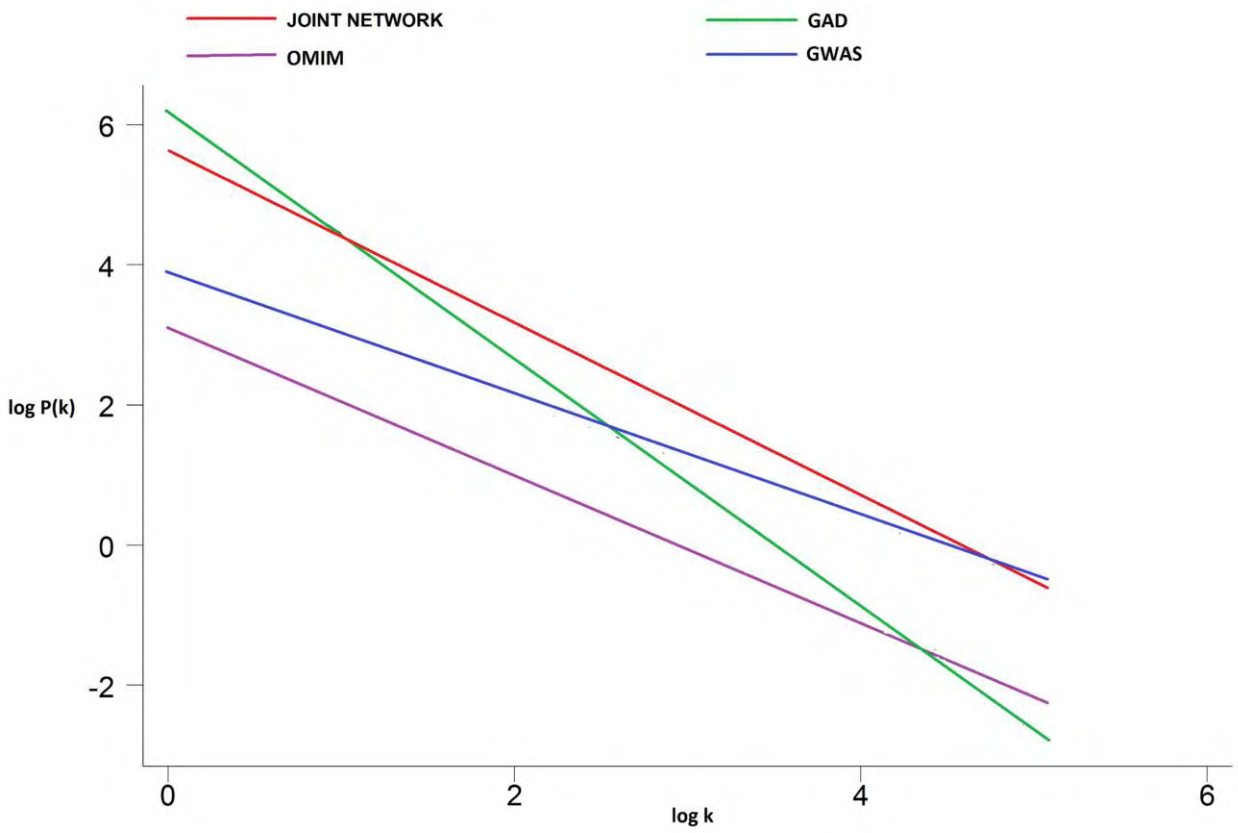
μικρές, υποστηρίζουν περαιτέρω το γεγονός ότι το δίκτυο είναι άνευ κλίμακας. Στην Εικόνα 5.9 παρουσιάζεται η κατανομή του βαθμού των κόμβων η οποία επιβεβαιώνει ότι ακολουθεί την κατανομή νόμου δύναμης. Το 78.8% των ασθενειών αλληλεπιδρά τουλάχιστον με μια μόνο ασθένεια και υπάρχουν 705 ασθένειες οι οποίες σχηματίζουν μια μεγάλη ομάδα με 3,956 αλληλεπιδράσεις οι οποίες αποτελούν το 56% των αλληλεπιδράσεων του συνολικού δικτύου (Εικόνα 5.8). Οι παρατηρήσεις αυτές καταλήγουν στη διαπίστωση ότι πολλές από τις παραπάνω ασθένειες εμφανίζουν ένα κοινό γενετικό υπόβαθρο. Αν οι ασθένειες είχαν διαφορετικό γενετικό υπόβαθρό τότε θα υπήρχε ένας μεγάλος αριθμός ασθενειών που θα αλληλεπιδρούσαν μόνο με ένα γονίδιο. Στις Εικόνες 5.5-5.8 εμφανίζεται μόνο ένας μικρός αριθμός ασθενειών οι οποίες αλληλεπιδρούν με ένα μόνο γονίδιο.

Ο μέσος βαθμός για κάθε κόμβο είναι ένα στατιστικό μέτρο που αντικατοπτρίζει την πυκνότητα του δικτύου και υπολογίζεται από τον αριθμό των ακμών/αριθμό των κόμβων του δικτύου. Ο μέσος βαθμός για κάθε κόμβο είναι: 2.2 στο δίκτυο της OMIM, 2.4 στο δίκτυο της GAD, 8.7 στο δίκτυο της GWAS και 5.7 στο συνολικό JOINT δίκτυο. Το δίκτυο της GWAS είναι το πιο πυκνό δίκτυο κι αυτό με το μεγαλύτερο αριθμό αλληλεπιδράσεων (ακμών). Το συνολικό δίκτυο είναι λιγότερο πυκνό σε σχέση με αυτό της GWAS και αυτό γιατί οι 3 βάσεις δεδομένων έχουν πολλές κοινές αλληλεπιδράσεις ασθενειών-ασθενειών και κατά τη δημιουργία του συνολικού δικτύου συνενώθηκαν, καταλήγοντας έτσι σε ένα λιγότερο πυκνό δίκτυο (Πίνακας 5.1).

Πίνακας 5.1. Στατιστικά μέτρα των δικτύων ασθενειών–ασθενειών.

	Κόμβοι	**Ακμές**	***γ***	**Μέσος Βαθμός Κόμβων**
OMIM	716	1573	1.77	2.2
GAD	69	168	1.02	2.4
GWAS	256	2230	0.83	8.7
JOINT NETWORK	998	5693	1.24	5.7

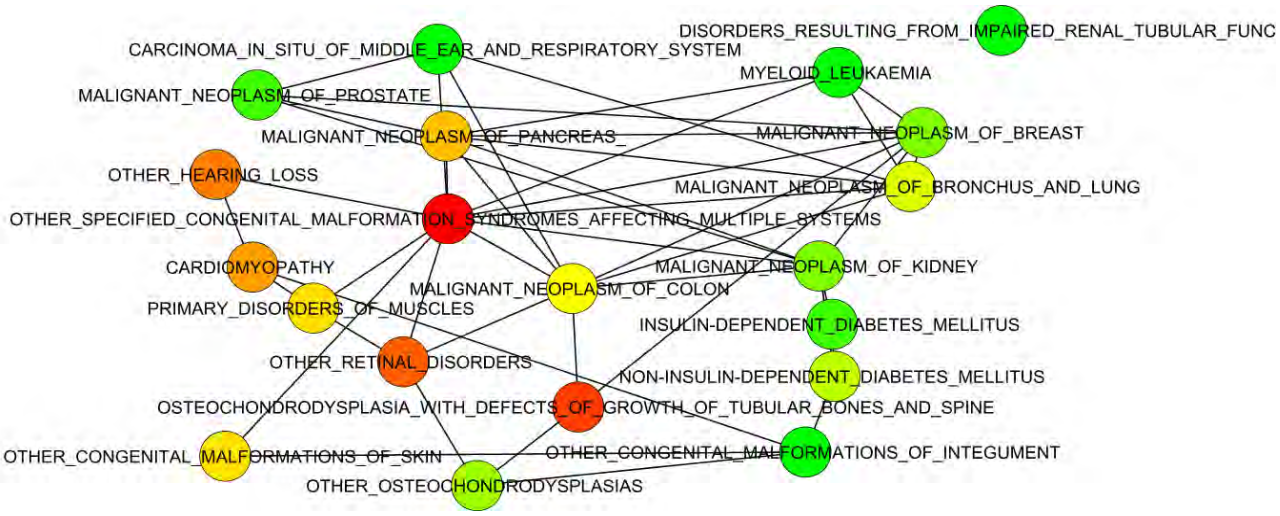
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



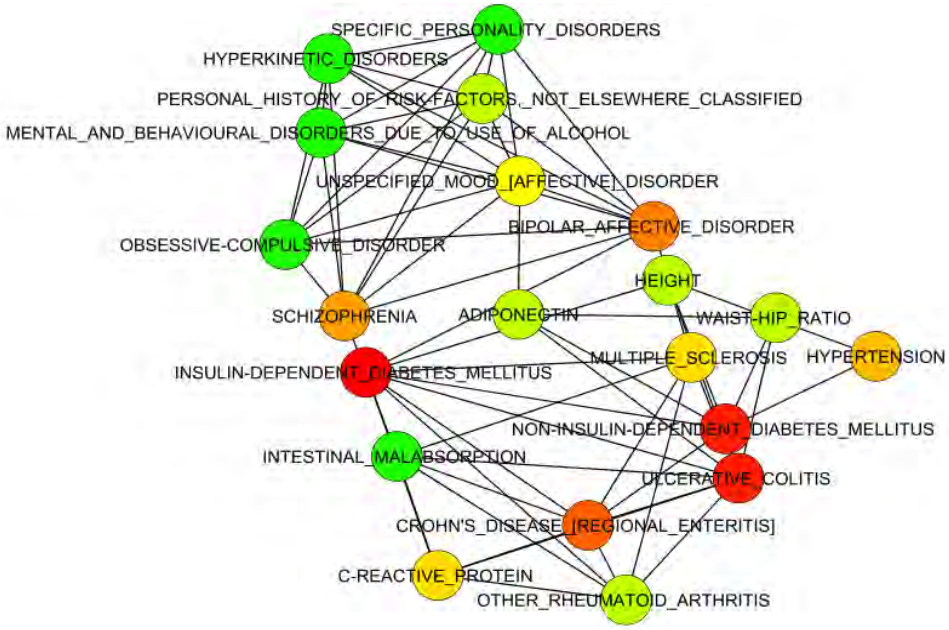
Εικόνα 5.9. Κατανομή του αριθμού των κόμβων σε λογαριθμική κλίμακα για τα δίκτυα ασθενειών-ασθενειών.

Όλοι οι κόμβοι των δικτύων ταξινομήθηκαν με βάση τον αριθμό των αλληλεπιδράσεων (ακμών) με σκοπό να βρεθούν οι κόμβοι (ασθένειες) με το μεγαλύτερο αριθμό αλληλεπιδράσεων. Στη συνέχεια, επιλέχθηκαν οι top-20 κόμβοι (Πίνακας 5.2), δηλαδή οι 20 ασθένειες οι οποίες αλληλεπιδρούν με μεγάλο αριθμό γονιδίων, και δημιουργήθηκαν τέσσερα υπο-δίκτυα με βάση αυτούς τους κόμβους στο πρόγραμμα cytoscape (Εικόνες 5.10-5.13). Σύμφωνα με το διάγραμμα Venn, η επικάλυψη μεταξύ των τεσσάρων διαφορετικών δικτύων είναι αξιοσημείωτα μικρή, με μία μόνο ασθένεια (μη ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης) να είναι κοινή και στα τέσσερα σύνολα δεδομένων (Εικόνα 5.14). Αυτές οι διαφορές οφείλονται κυρίως στο γεγονός ότι το περιεχόμενο των τριών βάσεων δεδομένων είναι αρκετά διαφορετικό.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

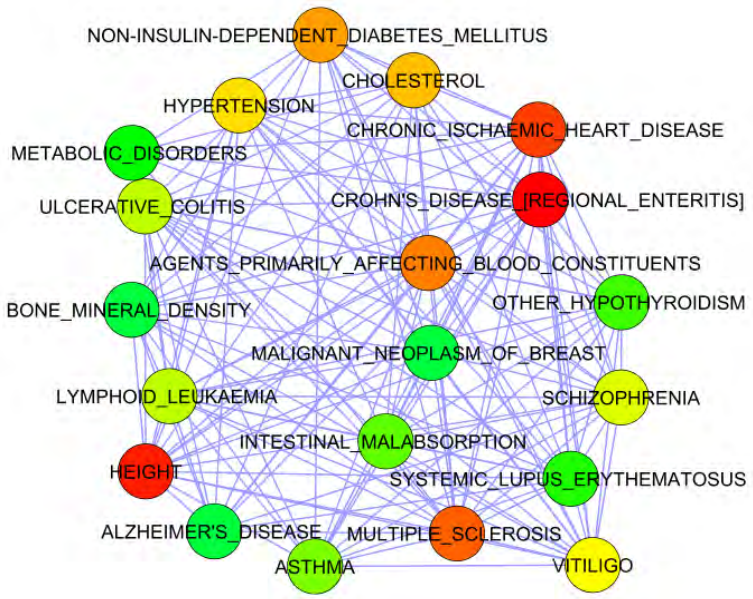


Εικόνα 5.10. Οι 20 ασθένειες με τη μεγαλύτερη συνδεσιμότητα στο δίκτυο ασθενειών-ασθενειών της OMIM.

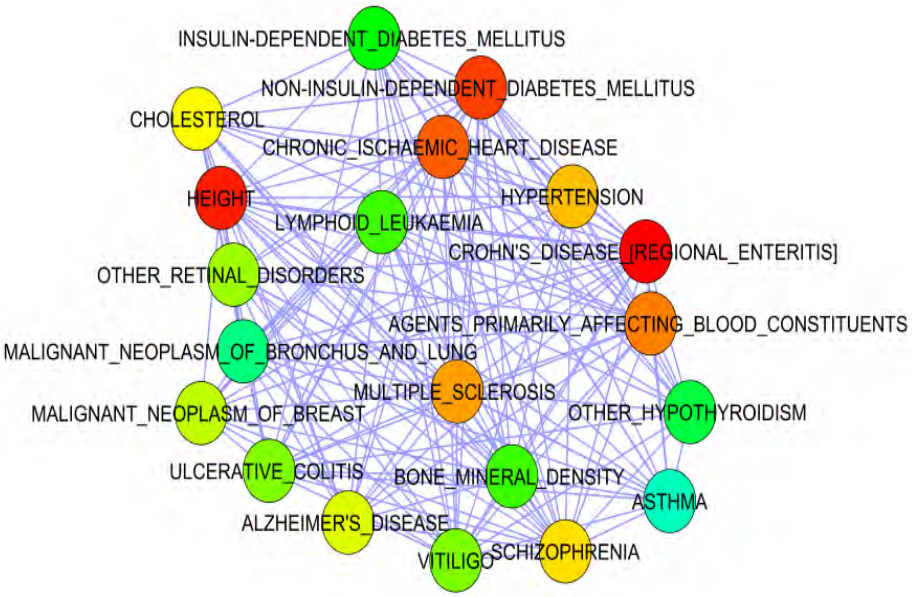


Εικόνα 5.11. Οι 20 ασθένειες με τη μεγαλύτερη συνδεσιμότητα στο δίκτυο ασθενειών-ασθενειών της GAD.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 5.12. Οι 20 ασθένειες με τη μεγαλύτερη συνδεσιμότητα στο δίκτυο ασθενειών-ασθενειών της GWAS.



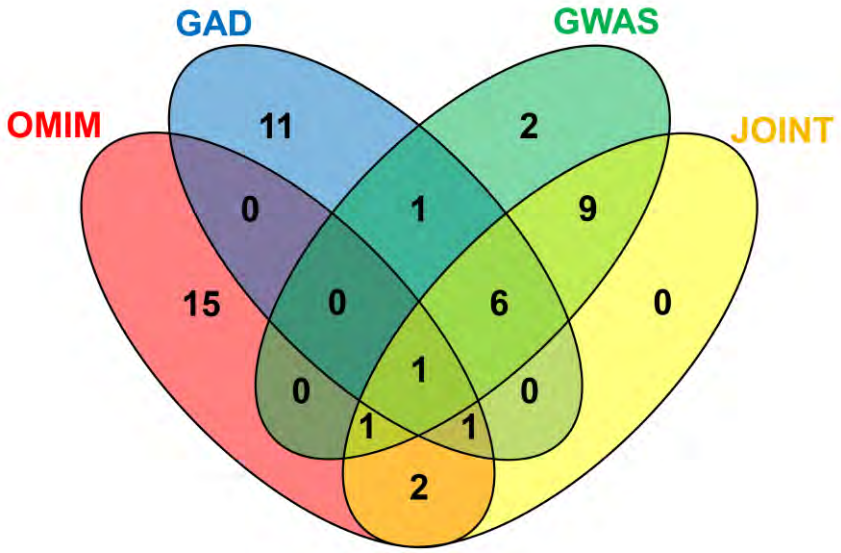
Εικόνα 5.13. Οι 20 ασθένειες με τη μεγαλύτερη συνδεσιμότητα στο δίκτυο ασθενειών-ασθενειών του συνολικού δικτύου.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 5.2. Οι 20 ασθένειες με τη μεγαλύτερη συνδεσιμότητα.

OMIM	**GAD**	**GWAS**	**Συνολικό Δίκτυο**
OTHER_SPECIFIED_CONGENITAL_MALFORMATION_SYNDROMES_AFFECTING_MULTIPLE_SYSTEMS	INSULIN-DEPENDENT_DIABETES_MELLITUS	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]
MALIGNANT_NEOPLASM_OF_COLON	ULCERATIVE_COLITIS	HEIGHT	HEIGHT
OSTEOCHONDRODYSPLASIA_WITH_DEFECTS_OF_GROWTH_OF_TUBULAR_BONES_AND_SPINE	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	CHRONIC_ISCHAEMIC_HEART_DISEASE	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS
OTHER_RETINAL_DISORDERS	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]	MULTIPLE_SCLEROSIS	CHRONIC_ISCHAEMIC_HEART_DISEASE
OTHER_HEARING_LOSS	BIPOLAR_AFFECTIVE_DISORDER	AGENTS_PRIMARILY_AFFECTING_BLOOD_CONSTITUENTS	AGENTS_PRIMARILY_AFFECTING_BLOOD_CONSTITUENTS
CARDIOMYOPATHY	SCHIZOPHRENIA	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	MULTIPLE_SCLEROSIS
MALIGNANT_NEOPLASM_OF_PANCREAS_	HYPERTENSION	CHOLESTEROL	HYPERTENSION
PRIMARY_DISORDERS_OF_MUSCLES	MULTIPLE_SCLEROSIS	HYPERTENSION	SCHIZOPHRENIA
OTHER_CONGENITAL_MALFORMATIONS_OF_SKIN	UNSPECIFIED_MOOD_[AFFECTIVE]_DISORDER	VITILIGO	CHOLESTEROL
MALIGNANT_NEOPLASM_OF_BRONCHUS_AND_LUNG	C-REACTIVE_PROTEIN	SCHIZOPHRENIA	ALZHEIMER'S_DISEASE
NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	OTHER_RHEUMATOID_ARTHRITIS	ULCERATIVE_COLITIS	MALIGNANT_NEOPLASM_OF_BREAST
OTHER_OSTEOCHONDRODYSPLASIAS	PERSONAL_HISTORY_OF_RISK-FACTORS	LYMPHOID_LEUKAEMIA	OTHER_RETINAL_DISORDERS
MALIGNANT_NEOPLASM_OF_KIDNEY	HEIGHT	ASTHMA	VITILIGO
MALIGNANT_NEOPLASM_OF_BREAST	WAIST-HIP_RATIO	INTESTINAL_MALABSORPTION	ULCERATIVE_COLITIS
INSULIN-DEPENDENT_DIABETES_MELLITUS	ADIPONECTIN	OTHER_HYPOTHYROIDISM	LYMPHOID_LEUKAEMIA
MALIGNANT_NEOPLASM_OF_PROSTATE	INTESTINAL_MALABSORPTION	SYSTEMIC_LUPUS_ERYTHEMATOSUS	BONE_MINERAL_DENSITY
CARCINOMA_IN_SITU_OF_MIDDLE_EAR_AND_RESPIRATORY_SYSTEM	SPECIFIC_PERSONALITY_DISORDERS	METABOLIC_DISORDERS	INSULIN-DEPENDENT_DIABETES_MELLITUS
MYELOID_LEUKAEMIA	OBSESSIVE-COMPULSIVE_DISORDER	BONE_MINERAL_DENSITY	OTHER_HYPOTHYROIDISM
OTHER_CONGENITAL_MALFORMATIONS_OF_INTEGUMENT	MENTAL_AND_BEHAVIOURAL_DISORDERS_DUE_TO_USE_OF_ALCOHOL	ALZHEIMER'S_DISEASE	MALIGNANT_NEOPLASM_OF_BRONCHUS_AND_LUNG
DISORDERS_RESULTING_FROM_IMPAIRED_RENAL_TUBULAR_FUNCTION	HYPERKINETIC_DISORDERS	MALIGNANT_NEOPLASM_OF_BREAST	ASTHMA

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 5.14. Το διάγραμμα Venn των top-20 κόμβων από τα τέσσερα δίκτυα ασθενειών-ασθενειών.

Στο δίκτυο της OMIM δεν παρατηρείται ιδιαίτερη συνδεσιμότητα και ένας κόμβος (ανάμεσα στους ισχυρά συνδεδεμένους) παραμένει ασύνδετος σε σχέση με τους υπόλοιπους. Αυτό πιθανότατα οφείλεται στο γεγονός ότι η OMIM περιλαμβάνει κυρίως μονογονιδιακές ασθένειες. Εν αντιθέσει με την OMIM, στη GWAS παρατηρείται μεγαλύτερη συνδεσιμότητα μεταξύ των ισχυρά συνδεδεμένων κόμβων, η οποία συνίσταται στο γεγονός ότι αυτή η βάση δεδομένων εμπεριέχει ένα μεγάλο αριθμό μονονουκλεοτιδικών πολυμορφισμών σε γονίδια που συνδέονται με μια κοινή ασθένεια. Επίσης, μεγάλη συνδεσιμότητα παρατηρείται στο συνολικό JOINT δίκτυο, όπως αναμενόταν, διότι είναι πιο ολοκληρωμένο εφόσον έχει ενσωματώσει δεδομένα και από τις τρεις διαφορετικές βάσεις δεδομένων.

Παρατηρώντας τις παραπάνω εικόνες, ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι φαινομενικά διαφορετικές ασθένειες - με διαφορετικούς φαινότυπους και κλινικές εκδηλώσεις- φαίνεται να είναι στενά συνδεδεμένες, υποδηλώνοντας ότι σχετίζονται με πολλά κοινά γονίδια. Ενδεικτικά, η σχιζοφρένεια, η σκλήρυνση κατά πλάκας και η νόσος Alzheimer, που αντιστοιχούν σε μια ψυχική διαταραχή, σε ένα αυτοάνοσο νόσημα και σε μια νευροεκφυλιστική νόσο, συνδέονται μεταξύ τους στο συνολικό δίκτυο και στο δίκτυο GWAS, υποδηλώνοντας ότι προκαλούνται από κοινά γονίδια. Παρόλα αυτά, οι ασθένειες αυτές παρουσιάζουν σημαντικές και ουσιαστικές ομοιότητες μεταξύ τους, όπως την ηλικία έναρξης των συμπτωμάτων στους ασθενείς, την εποχή του έτους την οποία οι ασθενείς έχουν γεννηθεί (στοιχείο, που μπορεί να είναι σημαντικό εάν ληφθεί υπόψη, για παράδειγμα, ότι μερικές ασθένειες μπορούν να προκληθούν σε διαφορετικές χρονικές περιόδους/καιρικές συνθήκες ή να

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ευδοκιμούν σε κρύο καιρό - ίσως λόγω καταστολής του ανοσοποιητικού συστήματος), ακόμη και τη συσχέτιση με την γεωγραφική θέση των ασθενών (Schwartz, 2011; Tremlett and Devonshire, 2006). Ωστόσο, η σχιζοφρένεια και η σκλήρυνση κατά πλάκας δεν συνδέονται στο δίκτυο GAD, παρόλο που αποτελούν κεντρικούς κόμβους˙ προφανώς λόγω της έλλειψης γονιδιακών δεδομένων. Επιπλέον, καμιά από τις τρεις ασθένειες δεν αποτελούν κεντρικούς κόμβους στο δίκτυο OMIM.

Επιπλέον, η υπέρταση, ένα καρδιαγγειακό νόσημα, και η λεμφοειδής λευχαιμία, ένα νεόπλασμα, συνδέονται στο δίκτυο GWAS και στο συνολικό δίκτυο. Σε προηγούμενη μελέτη έχει παρατηρηθεί υψηλή συχνότητα εμφάνισης της υπέρτασης σε παιδιά που πάσχουν από οξεία λεμφοβλαστική λευχαιμία (Attard-Montalto et al., 1994). Εν τούτοις, το γονιδιακό υπόβαθρο δεν ήταν γνωστό. Τα παραπάνω δίκτυα επιτρέπουν να ταυτοποιήσουμε τα κοινά γονίδια που συμμετέχουν στις δυο αυτές ασθένειες και δίνουν μια παραπάνω ένδειξη για την κατεύθυνση στην οποία πρέπει να κινηθούν οι μελλοντικές έρευνες.

Τέλος, το άσθμα, μια ασθένεια του αναπνευστικού συστήματος, συνδέεται με τη σχιζοφρένεια στο δίκτυο GWAS και στο συνολικό δίκτυο. Σε μια έρευνα που διεξήχθη στη Δανία παρατηρήθηκε μια συσχέτιση μεταξύ του άσθματος και του κίνδυνου ανάπτυξης σχιζοφρένειας (Pedersen et al., 2012). Όπως και στις προηγούμενες περιπτώσεις το γενετικό υπόβαθρο δεν είναι γνωστό.

Οι παραπάνω παρατηρήσεις επισημαίνουν την περαιτέρω ανάγκη για την ενσωμάτωση αλληλοσυμπληρωμένων δεδομένων από διαφορετικές πηγές στην δημιουργία δικτύων. Επίσης, με τη βοήθεια των δικτύων ασθενειών μπορούμε να εξάγουμε πολύ σημαντικά και χρήσιμα συμπεράσματα όσον αφορά τα γονίδια που ευθύνονται για φαινομενικά διαφορετικές σχετιζόμενες ασθένειες.

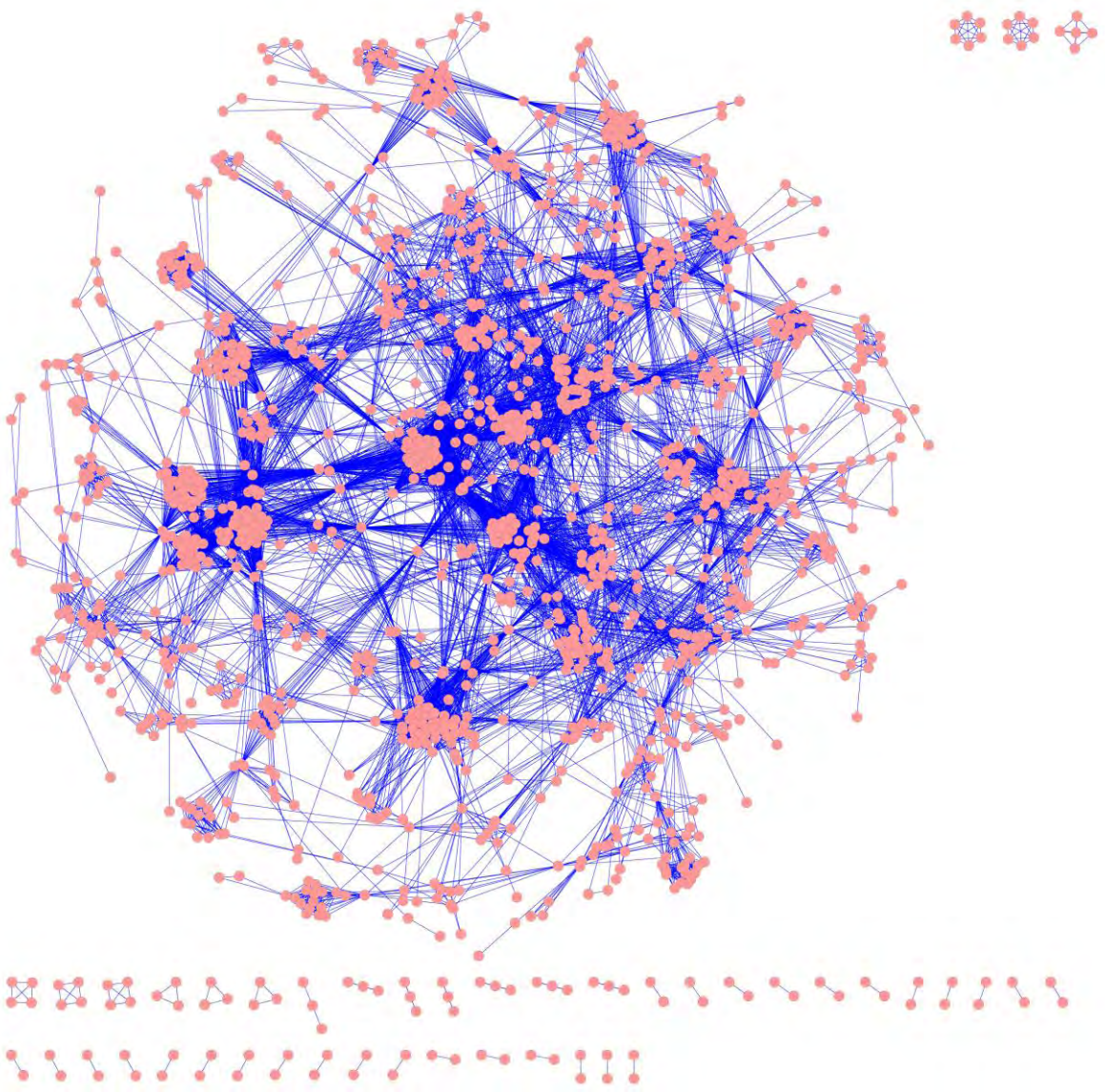
5.3.2 Δίκτυα γονιδίων-γονιδίων

Το δεύτερο σύνολο μονομερών δικτύων αποτελείται από τέσσερα δίκτυα γονιδίων, της OMIM, GAD, GWAS και το συνολικό JOINT δίκτυο (Εικόνες 5.15-5.18). Οι κόμβοι των δικτύων είναι τα γονίδια και οι μεταξύ τους συνδέσεις αποτελούν την ασθένεια με την οποία αλληλεπιδρούν. Το 92,9% των γονιδίων βρέθηκε να έχει παραπάνω από μια αλληλεπίδραση, κάτι το οποίο ενισχύει την άποψη ότι οι περισσότερες ασθένειες είναι προϊόν της δράσης αρκετών γονιδίων (multigenic).

Και στα τέσσερα δίκτυα γονιδίων-γονιδίων, υπάρχουν λίγοι κόμβοι οι οποίοι εμφανίζουν μεγάλο αριθμό συνδέσεων (κεντρικοί κόμβοι) (Εικόνες 5.15-5.18). Στην Εικόνα 5.19 παρουσιάζεται η κατανομή του αριθμού των κόμβων σε λογαριθμική κλίμακα για τα δίκτυα γονιδίων-γονιδίων που

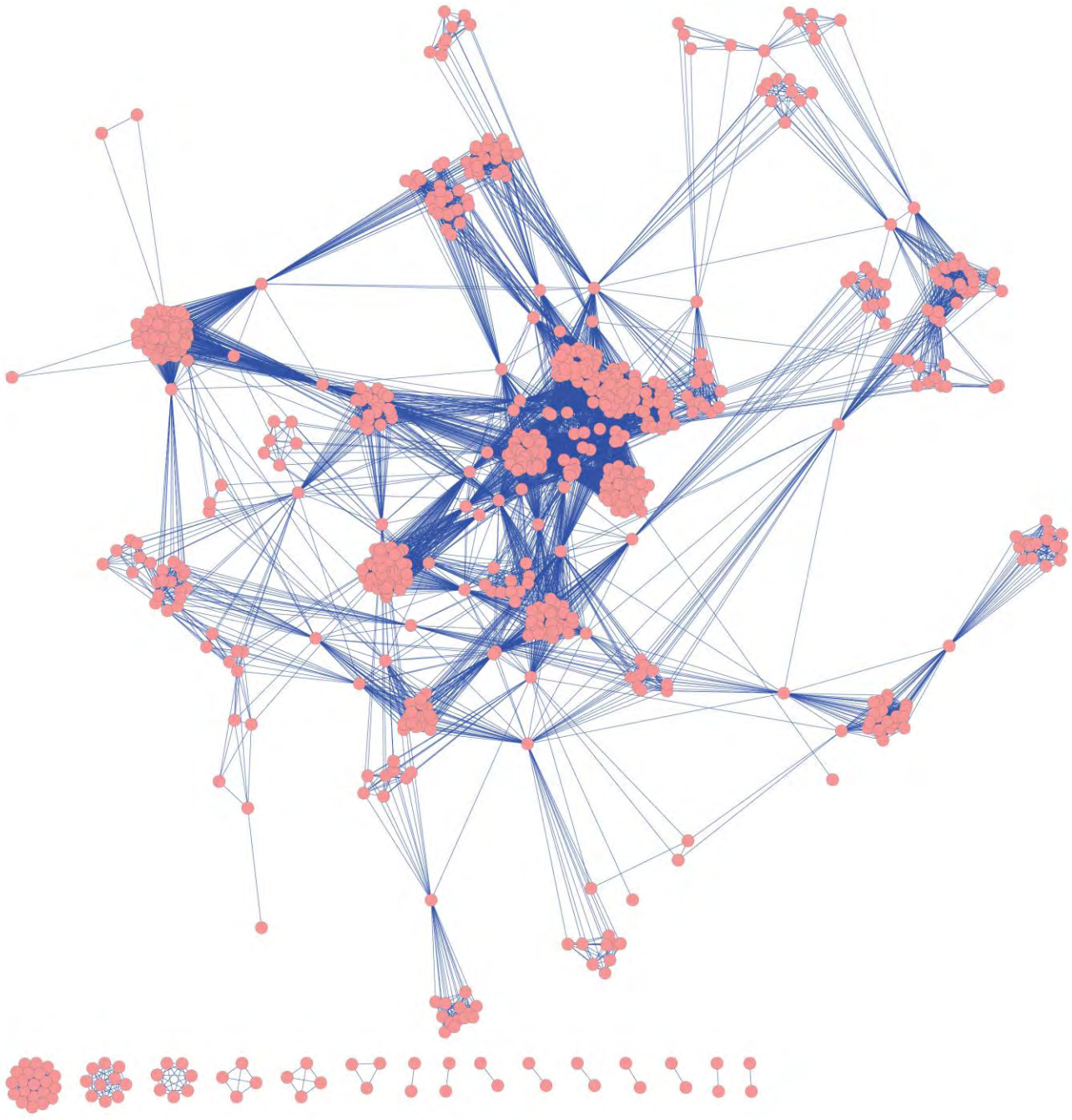
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

υποδεικνύει ότι και τα τέσσερα δίκτυα γονιδίων-γονιδίων χαρακτηρίζονται από μια κατανομή του βαθμού που ακολουθεί ένα νόμο δύναμης (power law) και προφανώς εμπίπτουν στην κατηγορία των δικτύων άνευ κλίμακας (scale free networks). Ο μέσος βαθμός των κόμβων είναι: 10.6 στην OMIM, 19.3 στη GAD, στην GWAS 40.3 και 39.4 στο συνολικό JOINT δίκτυο (Πίνακας 5.3). Το δίκτυο της GWAS είναι το πιο πυκνό (εικόνα 5.17, Πίνακας 5.3). Το συνολικό JOINT δίκτυο είναι ελάχιστα πιο αραιό από αυτό της GWAS και αυτό εξαιτίας κάποιας επικάλυψης μεταξύ των αλληλεπιδράσεων γονιδίων-γονιδίων. Όπως είναι αναμενόμενο τα δίκτυα γονιδίων είναι πιο πυκνά σε σχέση με τα δίκτυα των ασθενειών διότι είναι πιο μεγάλος ο αριθμός των γονιδίων που αλληλεπιδρούν με ένα σύνολο ασθενειών όπως αναφέρθηκε και παραπάνω.



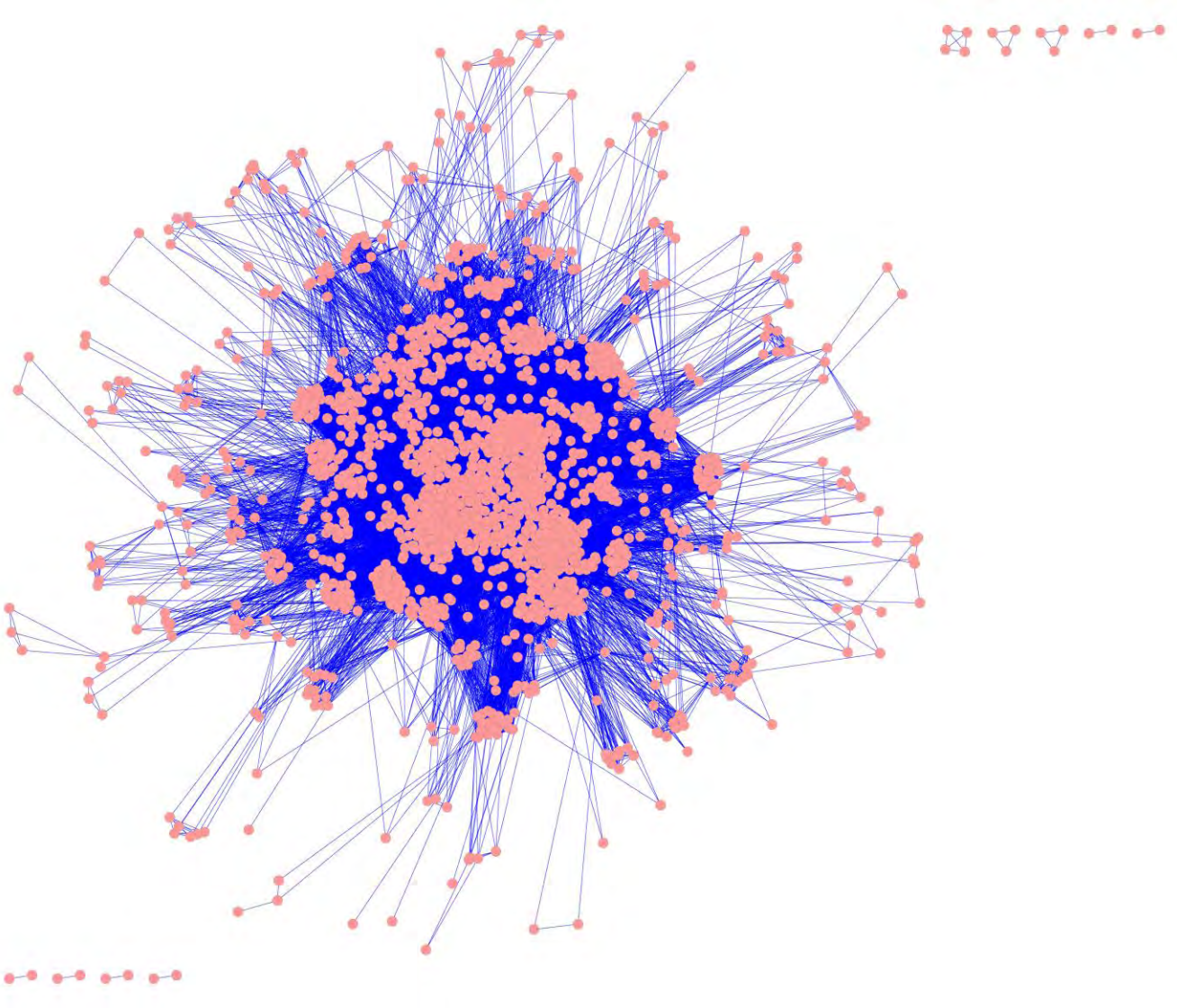
Εικόνα 5.15. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων της βάσης δεδομένων OMIM.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



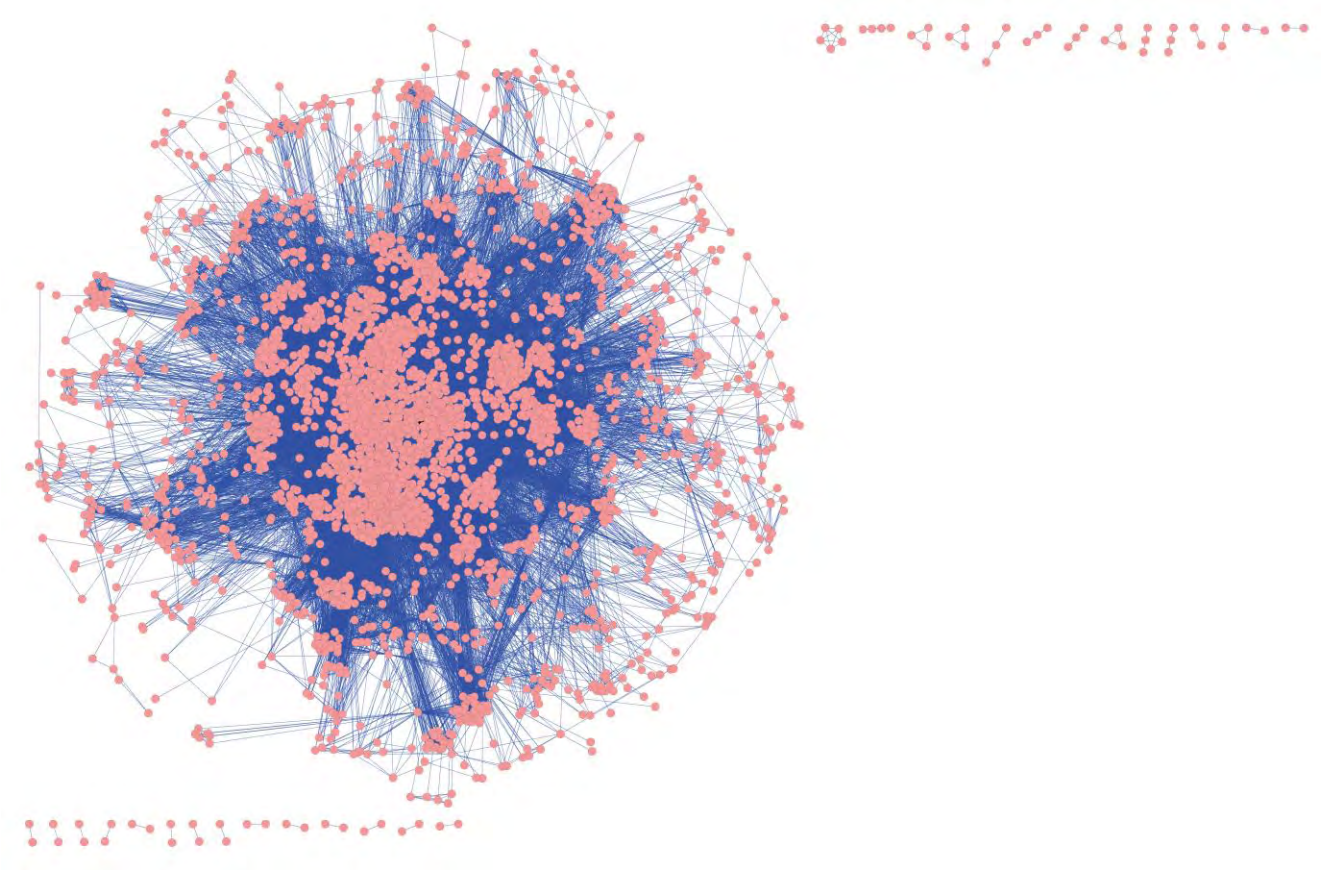
Εικόνα 5.16. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων της βάσης δεδομένων GAD.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 5.17. Το δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων της βάσης δεδομένων GWAS.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

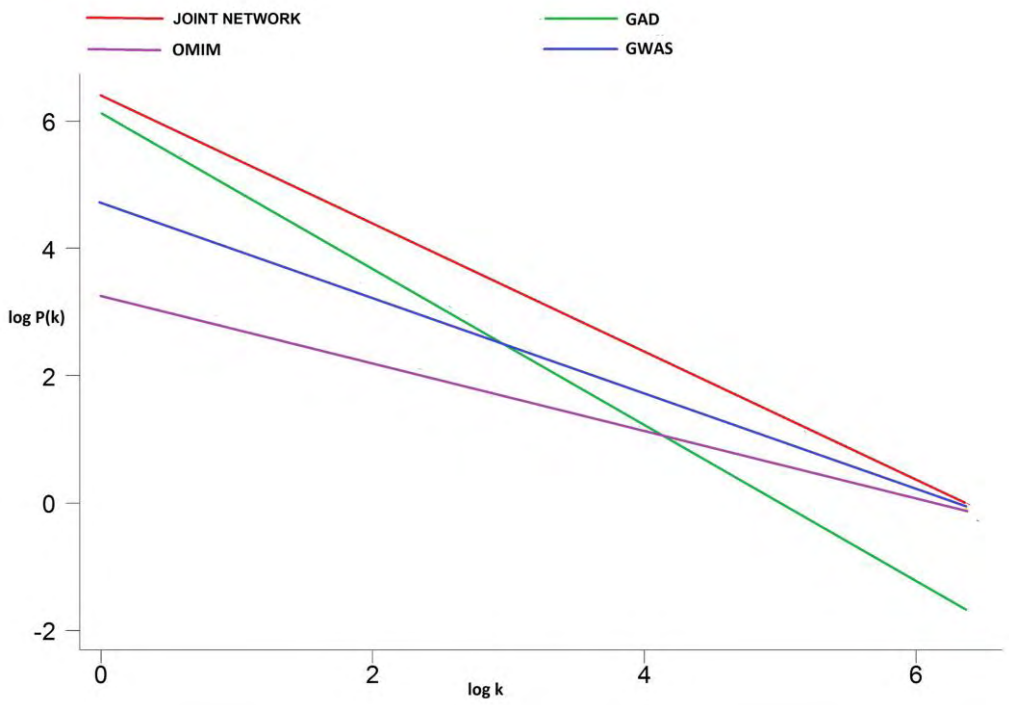


Εικόνα 5.18. Το συνολικό δίκτυο που περιλαμβάνει τις αλληλεπιδράσεις γονιδίων-γονιδίων.

Πίνακας 5.3. Στατιστικά μέτρα των δικτύων γονιδίων–γονιδίων.

	Κόμβοι	**Ακμές**	***γ***	**Μέσος Βαθμός Κόμβων**
OMIM	1543	16359	1.21	0.014
GAD	774	14903	0.55	0.050
GWAS	2005	80837	0.82	0.040
Συνολικό Δίκτυο	3580	140899	1.02	0.020

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



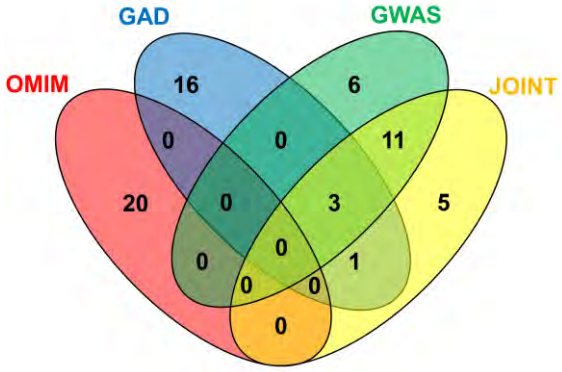
Εικόνα 5.19. Κατανομή του αριθμού των κόμβων σε λογαριθμική κλίμακα για τα δίκτυα γονιδίων-γονιδίων.

Στη συνέχεια για καθένα από τα δίκτυα υπολογίστηκε ο αριθμός των κόμβων με το μεγαλύτερο βαθμό (περισσότερες συνδέσεις-hubs) προκειμένου να βρεθούν ποια είναι αυτά που συσχετίζονται πιο πολύ με τις περισσότερες ασθένειες. Στον Πίνακα 5.4 παρουσιάζονται τα 20 γονίδια με το μεγαλύτερο αριθμό συνδέσεων στα δίκτυα των τριών βάσεων δεδομένων και του συνολικού. Παρατηρώντας τον πίνακα διαπιστώνεται ότι τα γονίδια διαφέρουν σημαντικά μεταξύ των τεσσάρων δικτύων. Πιο συγκεκριμένα, βάσει του διαγράμματος Venn, δεν υπάρχει ούτε ένας κοινός κεντρικός κόμβος μεταξύ των τεσσάρων δικτύων.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 5.4. Τα 20 γονίδια με τη μεγαλύτερη συνδεσιμότητα (Τα γονίδια με * αντιστοιχούν σε διαμεμβρανικές πρωτεΐνες).

OMIM	**GAD**	**GWAS**	**Συνολικό Δίκτυο**
SPINK5	IL2RA*	JAZF1	IRF8
KRAS	RPS14P1	HLA-DQA1*	BACH2
TRIM32	ST13P1	ADCY3*	MAP4K5
LMNA	PTPN2	C6ORF10*	LIPC
COL11A2	JAZF1	HFE*	NCKAP5
IL6	IRF8	MAP4K5	RASGRP1
NRAS	VEGFA	CDH13	CDH13
CFH	DAB2	BACH2	DNMT3A
TTC8	CENPW	HLA-DQB1*	KCNQ1*
CEP290	UBASH3A	SLC22A5*	HLA-DRA*
ARL6	C1ORF87	DNMT3A	SLC22A5*
FGFR2*	RPL34P26	KCNQ1*	SH2B3
SDHA	RPS3AP51	HLA-B*	ADCY3*
CREBBP	BACH2	UBE2L3	C6ORF10*
NF1	SMIM20*	IKZF1	NOTCH4*
ATP6V0A2*	SH2B3	IRF8	HLA-DQB1*
CAV3	PTPN22	LIPC	JAZF1
PTPN11	KIAA1109*	GPC5	HFE*
TTN	CTLA4*	ZMIZ1	BTNL2*
MYH7	PLCL1	THADA	HLA-DQA1*



Εικόνα 5.20. Το διάγραμμα Venn των top-20 κόμβων από τα τέσσερα δίκτυα γονιδίων-γονιδίων.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Στη συνέχεια, προκειμένου να επιβεβαιωθεί η ισχύς των αποτελεσμάτων, εφαρμόστηκαν διάφορες μέθοδοι τυχαιοποίησης και αντιμετάθεσης (permutation). Από την ανάλυση των αποτελεσμάτων των μεθόδων τυχαιοποίησης (είτε ανακατεύοντας τη μια στήλη του διμερούς γράφου είτε και τις δύο) παρατηρήθηκε ότι το 95% (19/20) των top-20 γονιδίων και των top-20 ασθενειών από τα αρχικά δίκτυα γονιδίων–γονιδίων και ασθενειών–ασθενειών βρέθηκε να κατατάσσεται μεταξύ των top 20 πιο συνδεδεμένων κόμβων στα τυχαία δίκτυα (Πίνακας 5.5-5.6). Ως εκ τούτου, ακόμη και μετά τη διαδικασία τυχαιοποίησης, τα τοπολογικά χαρακτηριστικά των δικτύων και οι ιδιότητες τους διατηρήθηκαν σταθερά. Για παράδειγμα, ένα γονίδιο που εμφανίζεται με μεγάλο βαθμό κόμβου στο αρχικό δίκτυο, εμφανίζεται επίσης με μεγάλο βαθμό κόμβου στα τυχαία δίκτυα, επειδή με βάση τα δεδομένα το γονίδιο αυτό εμπλέκεται σε πολλές ασθένειες. Το ίδιο αποτέλεσμα παρατηρήθηκε και όταν εφαρμόστηκε ο αλγόριθμος επανακαλωδίωσης (Gobbi et al., 2014). Ο αλγόριθμος επανακαλωδίωσης (rewiring) χρησιμοποιήθηκε επιπλέον για να ελεγχθεί το αν οι κατανομές του βαθμού των κόμβων των μονομερών δικτύων διατηρούνται σταθερές στα τυχαία δίκτυα ασθενειών- ασθενειών και γονιδίων–γονιδίων σε σχέση με τα αρχικά δίκτυα. Μετά από 100 προσομοιώσεις (simulations), οι τιμές της παραμέτρου γ από τα τυχαία δίκτυα δεν βρέθηκε να διαφέρουν σημαντικά ενώ ο αριθμός των αλληλεπιδράσεων (ακμών) παρέμενε σταθερός. Για παράδειγμα, η μέση τιμή της παραμέτρου γ για τα τυχαία δίκτυα ασθενειών βρέθηκε 1.314 (εύρος τιμών: 1.266-1.370), πολύ κοντά δηλαδή στην τιμή του πραγματικού δικτύου που είναι 1.386, ενώ η μέση τιμή της παραμέτρου γ για τα τυχαία δίκτυα γονιδίων βρέθηκε 1.129 (εύρος τιμών: 1.092-1.166), πολύ κοντά επίσης στην τιμή του πραγματικού δικτύου που είναι 1.143. Συνοψίζοντας, μετά από όλες τις διαδικασίες τυχαιοποίησης που πραγματοποιήθηκαν, οι top 20 πολύ συνδεδεμένοι κόμβοι του δικτύου των ασθενειών και των γονίδιων στα αρχικά δίκτυα βρέθηκαν να κατατάσσονται σταθερά μεταξύ των 20 κορυφαίων κόμβων στα τυχαιοποιημένα δίκτυα και οι τοπολογικές ιδιότητες των αρχικών δικτύων παρέμειναν ίδιες και στα τυχαία δίκτυα. Αυτό συμβαίνει επειδή το συνολικό JOINT δίκτυο είναι πολύ πυκνό και υπάρχει ένας μεγάλος αριθμός γονιδίων και ασθενειών που είναι ιδιαίτερα συνδεδεμένα. Ως εκ τούτου, ο αριθμός των γονιδίων που εμπλέκονται σε μια ασθένεια υπαγορεύει τη συνολική τοπολογία και την αρχιτεκτονική δικτύου.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Πίνακας 5.5. Οι top-20 πιο συνδεδεμένοι κόμβοι στο αρχικό δίκτυο γονιδίων - γονιδίων και στα τυχαία δίκτυα. Στην πρώτη στήλη απεικονίζονται τα top-20 γονίδια του αρχικού δικτύου. Στις επόμενες στήλες παρουσιάζονται τα top-20 γονίδια όπως βρέθηκαν από τα 100 τυχαία δίκτυα εφαρμόζοντας 3 διαφορετικές μεθόδους τυχαιοποίησης.

Top-20 (Αρχικό Δίκτυο)	**Top-20 στα τυχαία δίκτυα (με μετάθεση των ασθενειών)**	**Top-20 στα τυχαία δίκτυα (με μετάθεση και των γονιδίων και των ασθενειών)**	**Top-20 στα Rewired δίκτυα**
HLA-DQA1	HFE	HLA-DQA1	HLA-DQA1
JAZF1	HLA-DQA1	HLA-DRA	HLA-DQB1
C6ORF10	HLA-DQB1	HFE	ACE
SH2B3	APOE	ADCY3	JAZF1
HFE	JAZF1	IRF8	C6ORF10
LIPC	SH2B3	LIPC	SH2B3
NCKAP5	MAP4K5	NCKAP5	HFE
HLA-DQB1	LIPC	HLA-DQB1	LIPC
ADCY3	BTNL2	CDH13	NCKAP5
NOTCH4	HLA-DRA	BACH2	KCNQ1
CDH13	DNMT3A	C6ORF10	SLC22A5
BACH2	ADCY3	SH2B3	ADCY3
HLA-DRA	SLC22A5	KCNQ1	NOTCH4
RASGRP1	IRF8	SLC22A5	CDH13
DNMT3A	FTO	RASGRP1	BACH2
KCNQ1	C6ORF10	JAZF1	HLA-DRA
SLC22A5	NCKAP5	APOE	RASGRP1
IRF8	CDKN2B	MAP4K5	IRF8
MAP4K5	SLC6A4	NOTCH4	MAP4K5
BTNL2	KCNQ1	BTNL2	BTNL2

Πίνακας 5.6. Οι top-20 πιο συνδεδεμένοι κόμβοι στο αρχικό δίκτυο ασθενειών - ασθενειών και στα τυχαία δίκτυα. Στην πρώτη στήλη απεικονίζονται οι top-20 ασθένειες του αρχικού δικτύου. Στις επόμενες στήλες παρουσιάζονται οι top-20 ασθένειες όπως βρέθηκαν από τα 100 τυχαία δίκτυα, αφού εφαρμόστηκαν 3 διαφορετικές μέθοδοι τυχαιοποίησης.

Top-20 (Αρχικό Δίκτυο)	**Top-20 στα τυχαία δίκτυα (με μετάθεση των γονιδίων)**	**Top-20 στα τυχαία δίκτυα (με μετάθεση και των γονιδίων και των ασθενειών)**	**Top-20 στα Rewired δίκτυα**
CROHN'S_DISEASE_[REGIONAL_ENTERITIS]	HEIGHT	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]	HEIGHT
HEIGHT	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]	HEIGHT	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS
NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	SCHIZOPHRENIA	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	CHRONIC_ISCHAEMIC_HEART_DISEASE
CHRONIC_ISCHAEMIC_HEART_DISEASE	BIPOLAR_AFFECTIVE_DISORDER	CHRONIC_ISCHAEMIC_HEART_DISEASE	CROHN'S_DISEASE_[REGIONAL_ENTERITIS]
AGENTS_PRIMARILY_AFFECTING_BLOOD_CONSTITUENTS	MULTIPLE_SCLEROSIS	HYPERTENSION	MULTIPLE_SCLEROSIS
MULTIPLE_SCLEROSIS	NON-INSULIN-	CHOLESTEROL	HYPERTENSION

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

	DEPENDENT_DIABETES_MELLITUS		
HYPERTENSION	ULCERATIVE_COLITIS	ALZHEIMER'S_DISEASE	SCHIZOPHRENIA
SCHIZOPHRENIA	INSULIN-DEPENDENT_DIABETES_MELLITUS	SCHIZOPHRENIA	CHOLESTEROL
CHOLESTEROL	CHRONIC_ISCHAEMIC_HEART_DISEASE	VITILIGO	ALZHEIMER'S_DISEASE
ALZHEIMER'S_DISEASE	CHOLESTEROL	LYMPHOID_LEUKAEMIA	LYMPHOID_LEUKAEMIA
MALIGNANT_NEOPLASM_OF_BREAST	OTHER_RHEUMATOID_ARTHRITIS	INSULIN-DEPENDENT_DIABETES_MELLITUS	INSULIN-DEPENDENT_DIABETES_MELLITUS
OTHER_RETINAL_DISORDERS	OTHER_RETINAL_DISORDERS	OTHER_HYPOTHYROIDISM	ASTHMA
ULCERATIVE_COLITIS	ALZHEIMER'S_DISEASE	MALIGNANT_NEOPLASM_OF_BRONCHUS_AND_LUNG	BONE_MINERAL_DENSITY
VITILIGO	AGENTS_PRIMARILY_AFFECTING_BLOOD_CONSTITUENTS	AGENTS_PRIMARILY_AFFECTING_BLOOD_CONSTITUENTS	OTHER_HYPOTHYROIDISM
BONE_MINERAL_DENSITY	ASTHMA	MULTIPLE_SCLEROSIS	MALIGNANT_NEOPLASM_OF_BRONCHUS_AND_LUNG
LYMPHOID_LEUKAEMIA	OTHER_SPECIFIED_CONGENITAL_MALFORMATION_SYNDROMES_AFFECTING_MULTIPLE_SYSTEMS	MALIGNANT_NEOPLASM_OF_BREAST	OTHER_RETINAL_DISORDERS
INSULIN-DEPENDENT_DIABETES_MELLITUS	HYPERTENSION	OTHER_RETINAL_DISORDERS	MALIGNANT_NEOPLASM_OF_BREAST
OTHER_HYPOTHYROIDISM	PARKINSON'S_DISEASE	ULCERATIVE_COLITIS	ULCERATIVE_COLITIS
MALIGNANT_NEOPLASM_OF_BRONCHUS_AND_LUNG	OTHER_MENTAL_DISORDERS_DUE_TO_BRAIN_DAMAGE_AND_DYSFUNCTION_AND_TO_PHYSICAL_DISEASE	BONE_MINERAL_DENSITY	VITILIGO
ASTHMA	BONE_MINERAL_DENSITY	ASTHMA	MALIGNANT_NEOPLASM_OF_COLON

5.3.3 Μελέτη Περίπτωσης: ανάλυση αλληλεπιδράσεων διαμεμβρανικών πρωτεϊνών και διαμεμβρανικών υποδοχέων

Οι διαμεμβρανικές πρωτεΐνες, όπως είπαμε, αποτελούν μία μεγάλη και σημαντική οικογένεια των πρωτεϊνών. Διαδραματίζουν σημαντικό ρόλο σε πολλές κοινές ασθένειες όπως διαβήτης, υπέρταση, αρθρίτιδα, καρκίνος κ.ο.κ. Περισσότερο από το 50% του συνόλου των γνωστών φαρμάκων στοχεύουν διαμεμβρανικές πρωτεΐνες.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Προκειμένου να διερευνηθεί αυτή η συσχέτιση σε μεγάλη κλίμακα πραγματοποιήθηκε μια μελέτη περίπτωσης έτσι ώστε να βρεθεί ο τρόπος σύνδεσης των διαμεβρανικών πρωτεϊνών και των διαμεβρανικών υποδοχέων τόσο μεταξύ τους όσο και με άλλα γονίδια. Από τη βάση δεδομένων UNIPROT ανακτήθηκε ένα σύνολο διαμεμβρανικών πρωτεϊνών και το υποσύνολο αυτών, οι διαμεμβρανικοί υποδοχείς. Στη συνέχεια πραγματοποιήθηκε αναζήτηση των γονιδίων αυτών στο συνολικό JOINT δίκτυο και βρέθηκε ότι από τα 3854 μοναδικά γονίδια τα 1061 είναι γονίδια που κωδικοποιούν διαμεμβρανικές πρωτεΐνες (27,5%) και τα 289 είναι γονίδια που κωδικοποιούν διαμεμβρανικούς υποδοχείς (7,5%). Οι τιμές αυτές, είναι αναμενόμενες αν αναλογιστούμε ότι γενικά οι μεμβρανικές πρωτεΐνες απαντώνται σε όλους τους οργανισμούς σε ποσοστά που κυμαίνονται από 20-30%.

Στόχος της παρούσας μελέτης είναι να διερευνηθεί ο τρόπος «σύνδεσης», δηλαδή η αλληλεπίδραση των διεμεμβρανικών πρωτεϊνών και των διαμεμβρανικών υποδοχέων τόσο μεταξύ τους όσο και με άλλα γονίδια στο σχηματισμό της ασθένειας. Για το λόγο αυτό πραγματοποιήθηκε αναζήτηση στο συνολικό δίκτυο γονιδίων-γονιδίων για να βρεθεί πόσες διαμεμβρανικές πρωτεΐνες συνδέονται με μη διαμεμβρανικές πρωτεΐνες, πόσες διαμεβρανικές συνδέονται μεταξύ τους και αντίστοιχα η ίδια διερεύνηση πραγματοποιήθηκε και για τους διαμεμβρανικούς υποδοχείς. Στον Πίνακα 6.7 παρουσιάζονται τα αποτελέσματα σχετικά με τον αριθμό και το είδος των συσχετίσεων.

Οι συσχετίσεις μεταξύ των διαμεμβρανικών πρωτεϊνών αποτελούν το 7.8.% των συνολικών συσχετίσεων (των ακμών του δικτύου), ένα διόλου αμελητέο ποσοστό. Αξίζει να αναφερθεί, ότι απλά και μόνο λόγω τύχης, θα αναμέναμε οι διαμεβρανικές πρωτεΐνες να συμμετέχουν κατά 27,5%*27,5%=7.56% σε αλληλεπιδράσεις με άλλες διαμεμβρανικές πρωτεΐνες, και κατά 27,5%*2*72,5%=39.9% σε αλληλεπιδράσεις με μη-μεμβρανικές πρωτεΐνες (η παρατηρηθείσα τιμή είναι 39,2%, επίσης πολύ κοντά στην αναμενόμενη). Τέλος, οι μη-μεμβρανικές πρωτεΐνες, αλληλεπιδρούν μεταξύ τους σε ποσοστό 52,89%, τιμή επίσης πολύ κοντά στην αναμενόμενη (52,56%). Αυτά τα αποτελέσματα, αν και «αρνητικά», καθώς δεν δείχνουν κάποια απόκλιση από το αναμενόμενο, είναι ιδιαίτερα σημαντικά, καθώς δείχνουν ότι οι μεμβρανικές πρωτεΐνες δεν διαφέρουν σε τίποτα από τις μη-μεμβρανικές, όσον αφορά την παθογένεση και τον τρόπο με τον οποίο εμπλέκονται σε διάφορες ασθένειες. Αυτό, έρχεται σε συμφωνία με μια σειρά άλλα δεδομένα, όπως λ.χ. την ύπαρξη πολλών διαφορετικών λειτουργιών, αντίστοιχων με τις σφαιρικές πρωτεΐνες, που έχουν αποδοθεί σε διαμεμβρανικές πρωτεΐνες (ένζυμα, μεταφορείς, δομικές πρωτεΐνες, κλπ). Παρόλα αυτά, παραμένει η μεγάλη δυσκολία στον πειραματικό προσδιορισμό της δομής και της λειτουργίας τους, σαν συνέπεια

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

της δυσκολίας απομόνωσης από τις μεμβράνες και της δυσκολίας εύρεσης τρισδιάστατης δομής τους, και ίσως οι μελλοντικές μελέτες θα πρέπει να επικεντρωθούν σε αυτόν τον τομέα αλλά και σε υπολογιστικές προσπάθειες επίλυσής του (π.χ. με προγνωστικές μεθόδους κλπ).

Αξίζει να αναφερθεί τέλος, ότι στα 4 δίκτυα (OMIM, GAD, GWAS και το συνολικό), ανάμεσα στα 20 γονίδια με τις περισσότερες συνδέσεις, υπάρχουν 2, 4, 8 και 10 μεμβρανικές πρωτεΐνες αντίστοιχα. Τα δεδομένα αυτά, εκτός από τις διαφορές μεταξύ των βάσεων που αποκαλύπτουν, δείχνουν επίσης ότι συνολικά, οι μεμβρανικές πρωτεΐνες υπερ-εκπροσωπούνται ανάμεσα στις πρωτεΐνες με τις περισσότερες συνδέσεις (ο αναμενόμενος αριθμός θα ήταν 6). Αυτό σε αντίθεση με τα προηγούμενα που αφορούν το μέσο όρο των συνδέσεων ανά κατηγορία, δείχνει ότι ανάμεσα στους κεντρικούς κόμβους του δικτύου, υπάρχουν πολλές μεμβρανικές πρωτεΐνες, πράγμα που ενισχύει την πεποίθηση ότι παίζουν σημαντικό ρόλο στην ανάπτυξη πολλών και διαφορετικών ασθενειών.

Πίνακας 5.7. Συσχετίσεις μεταξύ διαφορετικών ζευγών γονιδίων στο συνολικό δίκτυο γονιδίων-γονιδίων. TM-PROTEIN είναι οι διαμεμβρανικές πρωτεΐνες που δεν είναι υποδοχείς. Κατά συνέπεια, ο συνολικός αριθμός των αλληλεπιδράσεων των διαμεμβρανικών πρωτεϊνών είναι το άθροισμα στη δεξιά στήλη.

Ζεύγη Γονιδίων		**Συσχετίσεις**	
NON-TM	NON-TM	74,530 (52,89%)	74,530
NON-TM	TM-PROTEIN	17,072 (12,12%)	55,273
NON-TM	TM-RECEPTOR	38,201 (27,11%)	
TM-PROTEIN	TM-PROTEIN	5,442 (3,86%)	11,096
TM-RECEPTOR	TM-RECEPTOR	1,175 (0,83%)	
TM-PROTEIN	TM-RECEPTOR	4,479 (3,18%)	
Συνολικός Αριθμός Συσχετίσεων Γονιδίων-Γονιδίων			140,899

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

5.4 Συμπεράσματα

Η παρούσα μελέτη παρέχει μια απεικόνιση των συσχετίσεων μεταξύ ασθενειών και γονιδίων. Τα στοιχεία που χρησιμοποιήθηκαν έχουν εξαχθεί από βάσεις δεδομένων και έχουν εφαρμοστεί αυστηρά κριτήρια σε μια προσπάθεια να αποκτηθούν αντιπροσωπευτικά και αξιόπιστα δεδομένα. Τονίστηκε η περαιτέρω ανάγκη για την ενσωμάτωση συμπληρωματικών δεδομένων από διαφορετικές πηγές για τη δημιουργία βιολογικών δικτύων, κάτι που έχει προταθεί και από προηγούμενες μελέτες (Bauer-Mehren et al., 2011). Η συνένωση και η μετα-ανάλυση πολύπλοκων βιολογικών δικτύων έχει ήδη προταθεί (Steele and Tucker, 2008) και ακολουθήθηκε και στη μελέτη αυτή. Έτσι, η διαδικασία που πραγματοποιήθηκε στην παρούσα μελέτη είναι απλή και εύκολη, αφού δημιουργήθηκαν τρία δίκτυα που προέρχονται από τρεις βάσεις δεδομένων που περιέχουν τις αλληλεπιδράσεις γονιδίων-ασθενειών με διαφορετικά χαρακτηριστικά. Αυτές οι βάσεις δεδομένων περιέχουν μη επεξεργασμένα δεδομένα και επιλέχθηκαν προκειμένου να αποφευχθεί το συστηματικό σφάλμα (bias) και να διατηρηθεί η αντιπροσωπευτικότητα των δεδομένων. Δεδομένου ότι το περιεχόμενο των βάσεων δεδομένων χαρακτηρίζεται πολλές φορές από θόρυβο και ασάφεια, στην παρούσα μελέτη έγινε προσπάθεια ελαχιστοποίησής τους με την ενοποίηση των ονομάτων των γονιδίων και ασθενειών των επιμέρους βάσεων δεδομένων με τη χρήση της HGNC και της ICD, αντίστοιχα. Με αυτόν τον τρόπο, αποδείχθηκε ότι υπήρχε σημαντική ετερογένεια στα ονόματα γονιδίων και των ασθενειών στις διάφορες βάσεις δεδομένων και αποφεύχθηκαν περιττές αλληλεπιδράσεις μεταξύ γονιδίων που είναι συνώνυμα ή επειδή η ασθένεια περιγράφεται με διαφορετικούς όρους.

Αξίζει να αναφερθεί ότι στη μελέτη που διεξήχθη από τον Goh και τους συνεργάτες του (Goh et al., 2007), ούτε οι όροι της ασθένειας, ούτε οι όροι του γονιδίου είχαν χρησιμοποιηθεί με κοινή ονοματολογία με αποτέλεσμα να υπάρξει ανομοιογένεια και λανθασμένη εκτίμηση των ευρημάτων. Στις τελευταίες μελέτες που διεξήχθησαν από τον Liu και τους συνεργάτες του (Liu et al., 2014) και τον Barrenas και τους συνεργάτες του (Barrenas et al., 2009), οι οποίοι συνένωσαν δεδομένα από διαφορετικές βάσεις δεδομένων (δυο από τις οποίες χρησιμοποιήθηκαν σε αυτή τη μελέτη), έγινε μια προσπάθεια να ομογενοποιηθεί η ονοματολογία των ασθενειών, αλλά όχι των γονιδίων. Συγκρίνοντας τα στοιχεία των κοινών βάσεων δεδομένων μεταξύ της παρούσας μελέτης και των μελετών που διεξήχθησαν από τον Liu και τους συνεργάτες του (Liu, Tseng et al. 2014) και τον Barrenas και τους συνεργάτες του (Barrenas, Chavali et al. 2009) προκύπτει ότι: στην παρούσα μελετη χρησιμοποιούνται 1897 γονίδια και 1059 ασθένειες από τη βάση δεδομένων OMIM, καθώς και 2018 γονίδια και 277

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ασθένειες από τη βάση δεδομένων GWAS. Στη μελέτη του Barrenas και των συνεργατών του (Barrenas et al., 2009) χρησιμοποιούνται 349 γονίδια και 54 πολύπλοκες ασθένειες, ένας πολύ μικρότερος αριθμός σε σύγκριση με εκείνον που αναλύεται στην παρούσα μελέτη, χωρίς να παρέχεται μια ενιαία ονοματολογία για τις καταχωρήσεις. Αντίθετα, στη μελέτη του Liu και των συνεργατών του (Liu et al., 2014), περιλαμβάνονταν 3749 γονίδια και 4457 ασθένειες από την OMIM, καθώς και 3397 γονίδια και 719 ασθένειες από τη βάση δεδομένων GWAS. Παρά το γεγονός ότι ο αριθμός των αλληλεπιδράσεων (εγγραφών) της μελέτης των Liu και των συνεργατών του (Liu et al., 2014) ήταν μεγαλύτερός σε σύγκριση με τον αριθμό που χρησιμοποιθηκε στην παρούσα μελέτη, οι καταχωρήσεις αυτές εξήχθησαν αυτόματα, χωρίς καμία προηγούμενη επιμέλεια και φιλτράρισμα, και ομογενοποιήθηκαν μόνο οι όροι των ασθενειών και όχι των γονιδίων.

Επιπλέον, οι βάσεις δεδομένων διαφέρουν επίσης σημαντικά όσον αφορά τις συσχετίσεις γονιδίου-ασθένειας. Η OMIM, όπως έχουμε ήδη συζητήσει, περιέχει υψηλής ποιότητας δεδομένα, που αφορούν κυρίως μονογονιδιακές ασθένειες και γονίδια υψηλής διεισδυτικότητας. Η GWAS περιέχει έγκυρες συσχετίσεις οι οποίες έχουν προκύψει και από επαναλήψιμες μελέτες, αλλά καλύπτει μόνο ένα μέρος των σημαντικών ασθενειών. Η GAD περιέχει δεδομένα που προκύπτουν από μελέτες γενετικών συσχετίσεων, δηλαδή από μελέτες που διερευνούν τη συσχέτιση ενός γονιδίου με μια ασθένεια με στατιστικό τρόπο, έτσι μπορεί να περιέχει πολλά ψευδή, μη επαναλήψιμα δεδομένα και αρνητικές συσχετίσεις των γονιδίων με τις ασθένειες. Έτσι, ανακτήθηκαν μόνο οι στατιστικά σημαντικές συσχετίσεις που απορρέουν από μελέτες μετα-ανάλυσης γενετικών δεδομένων από τη βάση αυτή. Λαμβάνοντας υπόψη όλα τα παραπάνω σχετικά με τα δεδομένα των ευρυγονιδιωματικών μελετών (GWAS) και των μελετών γενετικής συσχέτισης (GAS), φαίνεται ότι υπάρχει ανάγκη για ένα ενιαίο αποθετήριο το οποίο θα είναι διαθέσιμο στο ευρύ κοινό και θα περιέχει τις πιο αξιόπιστες (με βάση μια σειρά από κριτήρια) δημοσιευμένες γενετικές συσχετίσεις (Ioannidis et al., 2008).

Στη μελέτη αυτή, προτείνεται ότι με τη χρήση των δικτύων ασθενειών-ασθενειών, μπορεί να αντληθούν πολύ σημαντικά και χρήσιμα συμπεράσματα σχετικά με τη γενετική βάση των ασθενειών που φαινομενικά είναι διαφορετικές. Το δίκτυο γονιδίων-γονιδίων αποτελεί μια απεικόνιση της αρχιτεκτονικής των ασθενειών που εμφανίζονται στο άνθρωπο (human diseasome). Το δίκτυο αυτό θα μπορούσε να είναι ιδιαίτερα χρήσιμο για τον εντοπισμό των γενετικών παραγόντων που συμβάλλουν στην αιτιολογία, στην προδιάθεση (susceptibility) και στη συννοσηρότητα (comorbitity) των ασθενειών.

Λαμβάνοντας υπόψη όλα τα παραπάνω, τα αποτελέσματα αυτής της μελέτης μπορεί να έχουν σημαντική εφαρμογή στη διάγνωση, την πρόληψη και τη θεραπεία ασθενειών, καθώς και στο

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

σχεδιασμό φαρμάκων. Τα φαρμακολογικά δίκτυα αποτελούν ένα απαραίτητο εργαλείο για την ανάπτυξη νέων θεραπευτικών στρατηγικών (Berger and Iyengar, 2009; Hopkins, 2009; Wang et al., 2012b). Μια λεπτομερής κατανόηση του τρόπου με τον οποίο συνδέονται φαινομενικά διαφορετικές ασθένειες είναι απαραίτητη για την ανάπτυξη αποτελεσματικών θεραπειών. Μέχρι στιγμής, οι προσπάθειες έχουν κατευθυνθεί προς την αναζήτηση θεραπευτικών ουσιών που στοχεύουν τα γονίδια που συμμετέχουν σε συγκεκριμένες ασθένειες. Με τον τρόπο αυτό, οι ασθένειες που συνδέονται με γονίδια με γνωστή συσχέτιση αντιμετωπίζονται, ενώ άλλες ασθένειες με παρόμοιες κλινικές εκδηλώσεις δεν αντιμετωπίζονται (Kuhn et al., 2010; Nolan, 2007). Ως εκ τούτου, δύο ή περισσότερα φαινοτυπικά ανεξάρτητες ασθένειες με την ίδια γενετική προέλευση θα μπορούσαν να θεραπευθούν μέσω της δράσης ενός μόνο φαρμάκου. Οι σημαντικές συσχετίσεις ασθενειών–ασθενειών που ανιχνεύθηκαν στη μελέτη με κοινή γενετική βάση θα μπορούσαν να αποτελέσουν το θεμέλιο λίθο, στηριζόμενες στον οποίο θα κατευθυνθούν οι μελλοντικές προσπάθειες.

Τα προγράμματα που υλοποιήθηκαν, η μεθοδολογία που ακολουθήθηκε αλλά και τα αποτελέσματα που περιγράφηκαν στον παρόν Κεφάλαιο οδήγησαν σε 2 δημοσιεύσεις σε διεθνή επιστημονικά περιοδικά (Kontou et al., 2016c; Kontou et al., 2016d). Η εργασία στο *Gene* περιέχει τη βασική ανάλυση, ενώ η εργασία στο *Data in Brief* περιέχει λεπτομερή ανάλυση και παρουσίαση των προγραμμάτων και των δεδομένων. Επιπλέον, έχει δημιουργηθεί και ένα διαδικτυακό εργαλείο διαθέσιμο στο ευρύ κοινό με το οποίο ο χρήστης μπορεί να κάνει χρήση διαφόρων αλγορίθμων οπτικοποίησης και ομαδοποίησης δικτύων καθώς επίσης και να χρησιμοποιήσει τον αλγόριθμο μετατροπής των διμερών γράφων σε απλούς γράφους. Το εργαλείο είναι προσβάσιμο στη διεύθυνση και είναι ελεύθερο για ακαδημαϊκή χρήση http://www.compgen.org/tools/powerclust.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 6

Ανάλυση του Δικτύου Μεταγωγής Σήματος των Ανθρώπινων Υποδοχέων Συζευγμένων με G-πρωτεΐνες (GPCRS)

6.1 Εισαγωγή

Οι υποδοχείς συζευγμένοι με G-πρωτεΐνες (GPCRs) είναι η μεγαλύτερη ομάδα υποδοχέων στους ευκαρυωτικούς οργανισμούς (King et al., 2003) και αποτελούνται από επτά διαμεμβρανικές α-έλικες (Pierce et al., 2002). Το ανθρώπινο γονιδίωμα κωδικοποιεί περίπου 800 GPCRs, οι οποίοι μπορούν να ταξινομηθούν σε τέσσερις κύριες κατηγορίες με βάση την ομοιότητα στην αλληλουχία και στη λειτουργία τους (Attwood and Findlay, 1994; Kolakowski, 1994). Οι GPCRs αλληλεπιδρούν με ένα ευρύ φάσμα προσδετών (ligands) (πεπτίδια, πρωτεΐνες, προσταγλανδίνες, νευροδιαβιβαστές, φωτόνια, ορμόνες, ιόντα, φερορμόνες) (Gether, 2000; Okuno et al., 2008). Μετά από μια τέτοια αλληλεπίδραση, οι GPCRs υφίστανται μια διαμορφωτική αλλαγή (Kristiansen, 2004) και στη συνέχεια, το εξωκυτταρικό σήμα μεταδίδεται μέσα στο κύτταρο μέσω σύζευξης του με τις ετεροτριμερείς G-πρωτεΐνες (McCudden et al., 2005). Οι πρωτεΐνες αυτές σχηματίζουν ετεροτριμερή συμπλέγματα που αποτελούνται από τις G_{α}, G_{β} και G_{γ} υπομονάδες (Conklin and Bourne, 1993; Rens-Domiano and Hamm, 1995). Οι GPCRs αλληλεπιδρούν ειδικά με την G_{α} υπομονάδα των G-πρωτεϊνών μέσω ενδοκυτταρικού καρβοξυτελικού άκρο τους (Wong, 2003). Οι G-πρωτεΐνες στη συνέχεια ενεργοποιούν μόρια τελεστές (effectors), όπως τα κανάλια ιόντων και ενζύμων, και έτσι συμμετέχουν σε ποικίλες βιοχημικές και κυτταρικές αποκρίσεις (Cabrera-Vera et al., 2003; Trzaskowski et al., 2012). Τα μονοπάτια μεταγωγής σήματος των GPCR εμπλέκονται σε μια πληθώρα ασθενειών (Dorsam and Gutkind, 2007; Thompson et al., 2014; Vischer et al., 2014). Κατά συνέπεια, οι GPCRs έχουν τεράστια φαρμακολογική σημασία, με περίπου το 40% όλων των φαρμάκων στην αγορά να στοχεύουν αυτές τις πρωτεΐνες (Hopkins and Groom, 2002; Overington et al., 2006). Ως εκ τούτου, η αποσαφήνιση των μοριακών μηχανισμών του μονοπατιού της μεταγωγής σήματος των GPCR καθώς και ο τρόπος με τον οποίο το μονοπάτι συνδέεται με ασθένειες είναι υψίστης βιολογικής και φαρμακευτικής σημασίας.

Στόχος της παρούσας εργασίας είναι η μελέτη του συστήματος ligand-GPCR-G-protein-effector, στο εξής δίκτυο μεταγωγής σήματος των ανθρώπινων GPCR, έτσι ώστε να διερευνηθεί η πιθανή συσχέτιση της δράσης των υποδοχέων με τον κίνδυνο εμφάνισης ασθενειών. Ορισμένοι φαινότυποι ή παθολογικές εκδηλώσεις αποδίδονται σε γονίδια που αποτελούν λειτουργικές μονάδες, ενώ η απώλεια

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ενός μόνο γονιδίου μπορεί να διαταράξει τη λειτουργική δραστηριότητα της πλήρους μονάδας (Goh et al., 2007; Hartwell et al., 1999; Oti and Brunner, 2007). Για το σκοπό αυτό, χρησιμοποιήθηκαν λειτουργικά δίκτυα που περιλαμβάνουν τις σχέσεις των τεσσάρων τύπων μορίων του συστήματος σηματοδότησης των GPCR (ligand-GPCR-G-protein-effector). Ένα συνολικό δίκτυο κατασκευάστηκε από τις μεταξύ τους αλληλεπιδράσεις. Χρησιμοποιήθηκαν περαιτέρω πληροφορίες από δεδομένα γονιδιακής έκφρασης ανά ιστό (tissue-specific gene expression profiles) για την ενίσχυση της βιολογικής σημασίας της μελέτης μας. Δημιουργήθηκαν ειδικά-ανά-ιστό δίκτυα προκειμένου να ανιχνευθεί αν υπάρχουν λειτουργικές μονάδες που εκφράζονται σε συγκεκριμένους ιστούς. Ένας άλλος στόχος αυτής της μελέτης ήταν να διερευνήσει το πώς τα προφίλ έκφρασης των ειδικά-ανά-ιστό γονιδίων που σχετίζονται με ασθένεια συσχετίζονται επιπλέον και με το σχετικό κίνδυνο ανάπτυξης της ασθένειας. Επίσης, ανιχνευθήκαν χαρακτηριστικά μονοπάτια που εμφανίζονται σε συγκεκριμένες παθολογικές καταστάσεις. Τέλος, εξετάστηκαν αλληλεπιδράσεις των μορίων του συστήματος σηματοδότησης GPCR με άλλες πρωτεΐνες αλλά και με γνωστά φάρμακα.

6.2 Μέθοδοι

6.2.1 Συλλογή Δεδομένων

Μοριακές αλληλεπιδράσεις

Όλοι οι ανθρώπινοι GPCRs, G-πρωτεΐνες (Gα υπομονάδα), οι τελεστές, καθώς και οι αντίστοιχες αλληλεπιδράσεις τους, συλλέχθηκαν από τη βάση δεδομένων Human-gpDB (Satagopam et al., 2010) (http://bioinformatics.biol.uoa.gr/human_gpdb) στις 15 Ιουλίου, 2014. Η κατάταξη όλων των γνωστών ανθρώπινων GPCRs ανακτήθηκε από τις βάσεις δεδομένων IUPHAR/BPS Guide to PHARMACOLOGY (Southan et al., 2016) και Human-gpDB (Satagopam et al., 2010). Επιπλέον, οι πεπτιδικοί προσδέτες (peptide ligands) των ανθρώπινων GPCRs συλλέχθηκαν από τη βάση δεδομένων UniProt (UniProt_Consortium, 2012). Οι αλληλεπιδράσεις των πεπτιδικών προσδετών με τους αντίστοιχους GPCRs ανακτήθηκαν από τη βάση δεδομένων βιβλιογραφίας MEDLINE μετά από εκτεταμένη βιβλιογραφική έρευνα.

Δεδομένα γονιδιακής έκφρασης

Τα προφίλ γονιδιακής έκφρασης από 11 φυσιολογικούς ανθρώπινους ιστούς (υποθάλαμος, σπλήνα, ωοθήκες, πνεύμονας, ήπαρ, νεφρά, καρδιά, παχύ έντερο, λιπώδης ιστός, όρχεις και σκελετικό μυ)

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ανακτήθηκαν από τη βάση δεδομένων RNA-Seq Atlas (Krupp et al., 2012) (http://medicalgenomics.org/rna_seq_atlas) στις 24 Οκτωβρίου 2014 για καθένα από τα γονίδια των προσδετών (ligangs), των υποδοχέων GPCRs, των G-πρωτεϊνών και των μορίων τελεστές (effectors).

Συσχετίσεις γονιδίων-ασθενειών

Η συσχέτιση των γονιδίων των προσδετών (ligangs), των υποδοχέων GPCRs, των G-πρωτεϊνών και των μορίων τελεστές (effectors) με ασθένειες ανακτήθηκε από τις βάσεις δεδομένων Genetic Association Database (GAD) (Becker et al., 2004), Genome Wide Association Studies (GWAS) (Welter et al., 2014) και Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015). Προκειμένου να διατηρηθεί μια κοινή ονοματολογία μεταξύ των ασθενειών και των γονιδίων μεταξύ των τριών βάσεων δεδομένων χρησιμοποιήθηκε το Διεθνές Σύστημα Ταξινόμησης Ασθενικών (ICD) για τη μετατροπή των ονομάτων των ασθενειών και το σύστημα HGNC (HUGO Gene Nomenclature Committee) (Gray et al., 2015) για τη μετατροπή των ονομάτων των γονιδίων. Από την ενσωμάτωση αυτών των στοιχείων από τις τρεις διαφορετικές πηγές κατασκευάστηκε ένα δίκτυο γονιδίων-ασθενειών το οποίο περιλαμβάνει συσχετίσεις γονιδίων με μονογονιδιακές αλλά και με πολυπαραγοντικές ασθένειες, το οποίο αναλύθηκε στο προηγούμενο Κεφάλαιο (Kontou et al., 2016d; Pantavou et al., 2016).

Συσχέτιση ασθενειών με ιστούς

Οι πληροφορίες σχετικά με τις ασθένειες που εμπλέκονται σε συγκεκριμένες ιστούς ανακτήθηκαν από τη βάση δεδομένων PhenoDigm (Oellrich et al., 2014; Smedley et al., 2013) η οποία περιλαμβάνει τις συσχετίσεις ιστού-ασθένειας και από εκτενή αναζήτηση στη βιβλιογραφία προκειμένου να βρεθούν οι ασθένειες που εκφράζονται στους 11 διαφορετικούς ιστούς που μελετώνται.

Φάρμακα-στόχοι σε μόρια του δικτύου των GPCR

Για κάθε τύπο μορίου (προσδέτες, GPCRs, G-πρωτεΐνες, τελεστές) πραγματοποιήθηκε αναζήτηση στις βάσεις δεδομένων UniProt (UniProt, 2015) και DrugBank (Law et al., 2014) προκειμένου να υπολογιστεί ο συνολικός αριθμός των αλληλεπιδρώντων φαρμάκων.

Αλληλεπιδράσεις των μορίων του δικτύου των GPCR

Οι αλληλεπιδράσεις μεταξύ των τεσσάρων τύπων μορίων του δικτύου, όπως επίσης και οι αντίστοιχες αλληλεπιδράσεις τους με άλλες πρωτεΐνες οι οποίες δεν βρίσκονται στο δίκτυο της μεταγωγής σήματος

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

των GPCR ανακτήθηκαν από τη βάση δεδομένων Intact (Orchard et al., 2014) η οποία περιέχει αλληλεπιδράσεις πρωτεϊνών -πρωτεϊνών (PPIs).

6.2.2 Οπτικοποίηση των δεδομένων γονιδιακής έκφρασης

Τα προφίλ γονιδιακής έκφρασης από τη RNA-Seq Atlas (Krupp et al., 2012) για κάθε ιστό απεικονίστηκαν γραφικά με το λογισμικό HeatmapGenerator (Khomtchouk et al., 2014).

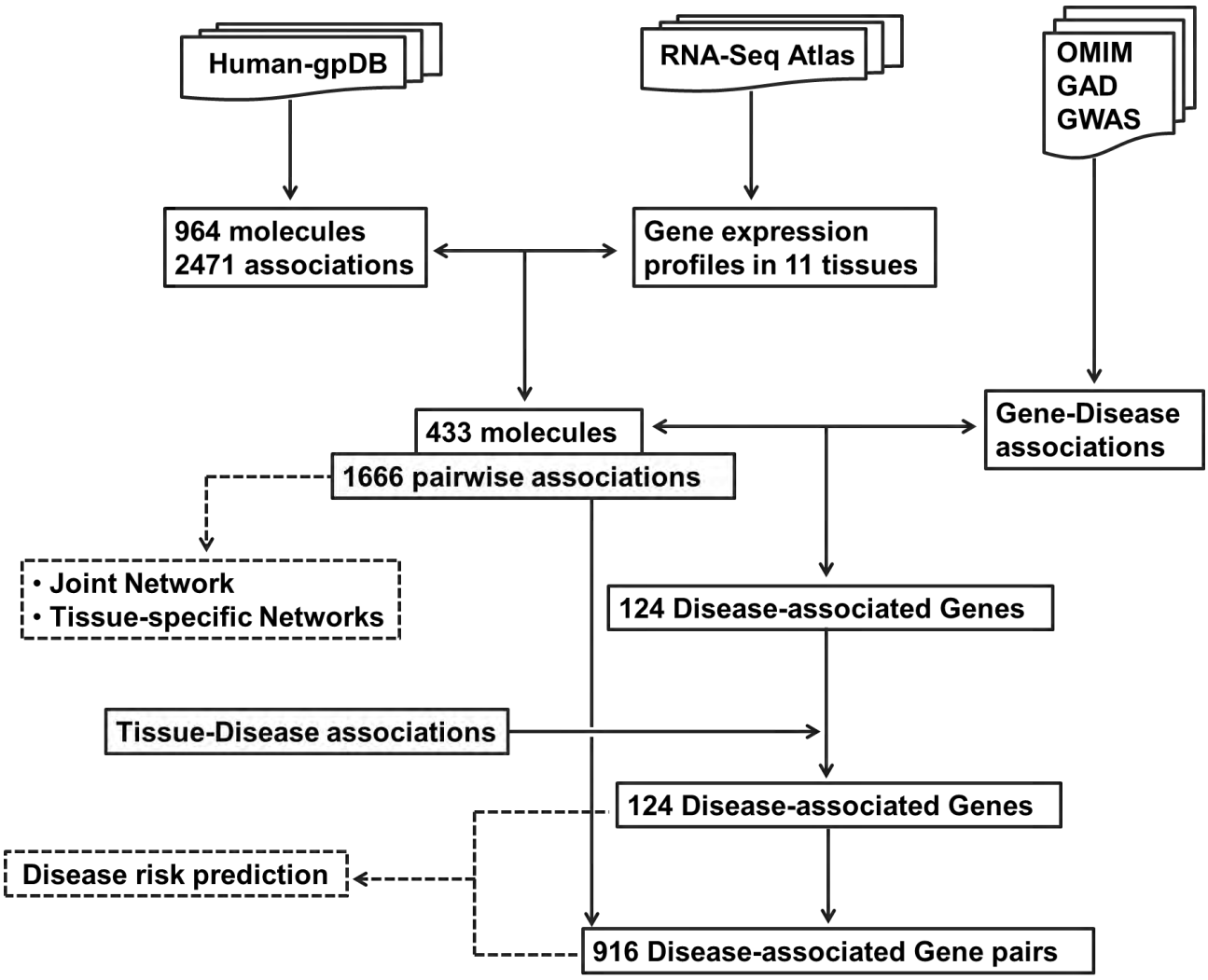
6.2.3 Στατιστική ανάλυση και ανάλυση δικτύων

Το λογισμικό Cytoscape v.3.2.1 (http://www.cytoscape.org) (Shannon et al., 2003) χρησιμοποιήθηκε για τη στατιστική ανάλυση, επεξεργασία και οπτικοποίηση των δεδομένων του δικτύου. Η ανάλυση του δικτύου επικεντρώθηκε στις τοπολογικές ιδιότητες του δικτύου μεταγωγής σήματος των ανθρώπινων GPCRs.

6.2.4 Κατασκευή μοντέλου εκτίμησης του κινδύνου εμφάνισης της νόσου

Ένα μοντέλο λογιστικής παλινδρόμησης κατασκευάστηκε για την πρόβλεψη του σχετικού κινδύνου ανάπτυξης μιας ασθένειας για την περίπτωση των κατά ζεύγη αλληλεπιδράσεων των συστατικών του συστήματος σηματοδότησης των GPCR (δηλαδή για τα ζεύγη προσδέτης-GPCR, GPCR-G-πρωτεΐνη και G-πρωτεΐνη-τελεστής). Υπολογίστηκαν τα τυπικά σφάλματα που αντιπροσωπεύουν τον αριθμό των κατά ζεύγη αλληλεπιδράσεων. Ειδικότερα, στο μοντέλο χρησιμοποιήθηκε η πληροφορία για τον τύπο του αλληλεπιδρώντων μορίων, το είδος του ιστού, τη λογαριθμική μέση τιμή της γονιδιακής έκφρασης, το άθροισμα του αριθμού των φαρμάκων, το άθροισμα του συνολικού αριθμού των αλληλεπιδράσεων και το άθροισμα των εξερχόμενων ακμών που προέρχονται από έναν κόμβο και των εισερχόμενων ακμών πάνω σε έναν κόμβο ανά ιστό των εμπλεκόμενων γονιδίων. Όλες οι στατιστικές αναλύσεις πραγματοποιήθηκαν με το στατιστικό λογισμικό Stata 13 (StataCorp.2013). Στη συνέχεια, η πληροφορία σχετικά με τη συσχέτιση των γονιδίων με ασθένειες αξιοποιήθηκε έτσι ώστε να βρεθούν, αν υπάρχουν, μονοπάτια γονιδίων τα οποία να εμφανίζουν και στους 4 κόμβους τους κοινές ασθένειες και να διερευνηθεί περαιτέρω η αιτιολογία των ασθενειών αυτών.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.1. Η διαδικασία συλλογής και επεξεργασίας των δεδομένων του δικτύου μεταγωγής σήματος των ανθρώπινων GPCRs.

6.3 Αποτελέσματα

Τα μόρια του συστήματος σηματοδότησης των GPCR (ligand-GPCR-G-protein-effector) αλληλεπιδρούν μεταξύ τους μέσω ζευγών αλληλεπιδράσεων/συσχετίσεων. Συνολικά, 964 μόρια και 2471 ζεύγη των μεταξύ τους αλληλεπιδράσεων ανακτήθηκαν από τη Humam-gpDB. Προκειμένου να βρεθεί η γονιδιακή έκφραση στους ιστούς των παραπάνω μορίων πραγματοποιήθηκε αναζήτηση στη βάση δεδομένων RNA-Seq Atlas και ανακτήθηκαν 483 μόρια με τις αντίστοιχες αλληλεπιδράσεις τους. Δεδομένου ότι σε 32 από τα 483 μόρια η τιμή της έκφρασης ήταν ίση με μηδέν σε όλους τους ιστούς, οι καταχωρήσεις αυτές αφαιρέθηκαν από τα επόμενα στάδια αυτής της μελέτης. Επιπλέον, αφαιρέθηκαν

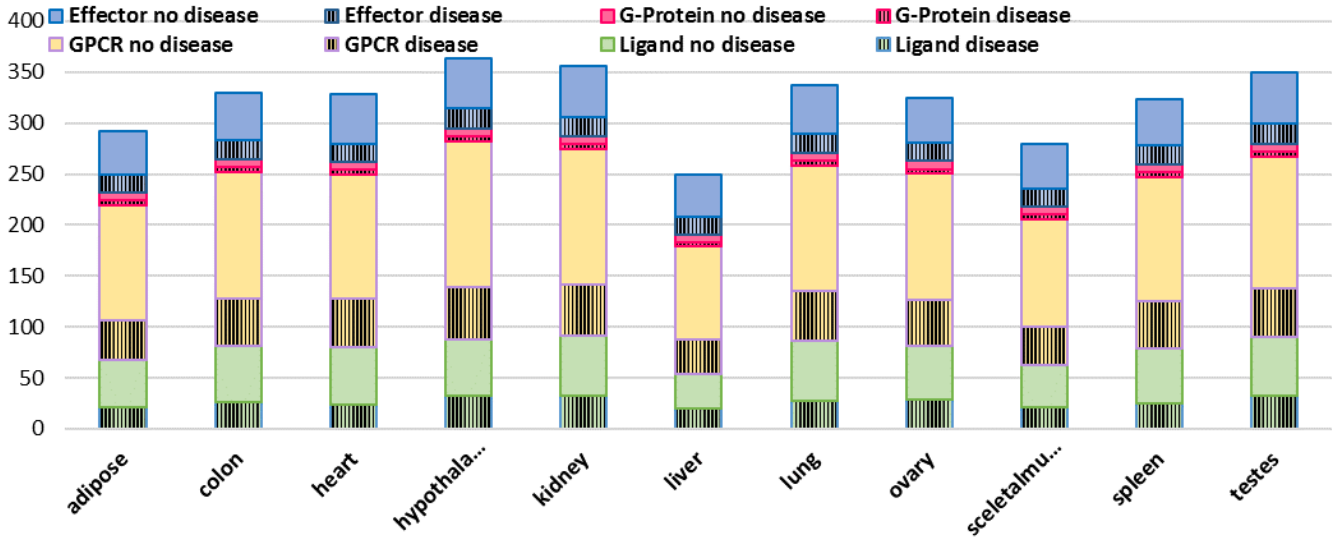
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

18 μόρια που συνδέονται αποκλειστικά με αυτά τα 32 μόρια. Ως εκ τούτου, ένα σύνολο 433 μορίων και 1666 ζευγών αλληλεπιδράσεων αναλύθηκαν σε αυτήν τη μελέτη. Η κατανομή των 433 μορίων ανά ιστό φαίνεται διαγραμματικά στην Εικόνα 6.2. Από τα 433 μόρια, πάνω από τα μισά αντιπροσωπεύουν GPCRs (Εικόνα 6.2). Αντίθετα, οι G-πρωτεΐνες υποεκπροσωπούνται, με μόνο 13 μόρια.

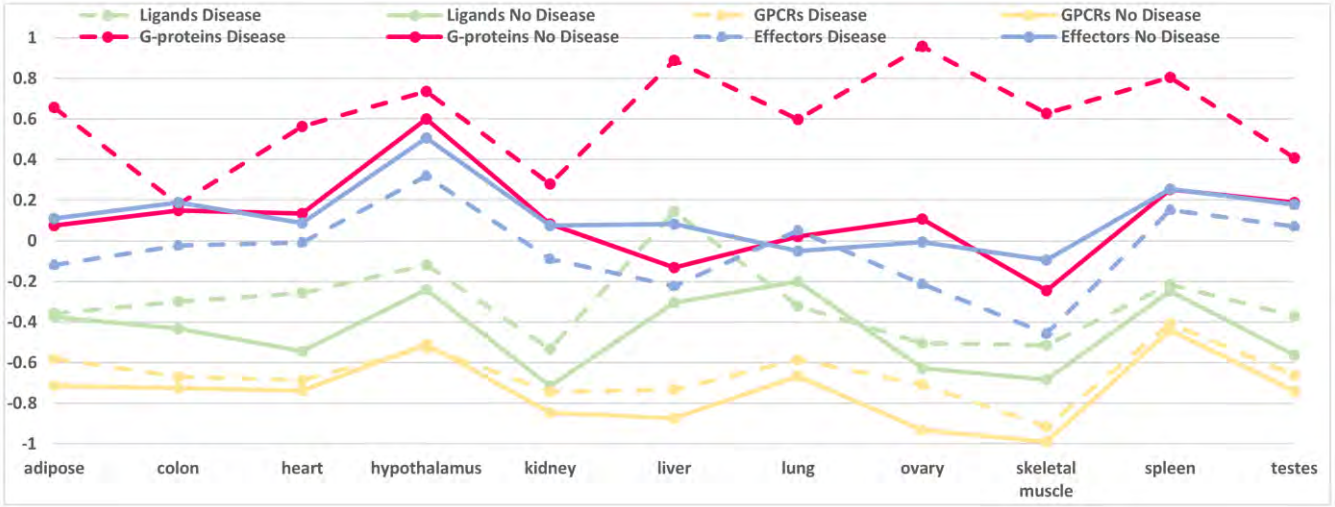
6.3.1 Γονιδιακή έκφραση ειδική ανά ιστό

Οι γονιδιακή έκφραση των 433 γονιδίων σε 11 υγιείς ιστούς ανακτήθηκε από την RNA-Seq Atlas. Επτά γονίδια (AGT (αγγειοτενσινογόνο), ANXA1 (αννεξίνη A1), APP (β-αμυλοειδής πρόδρομη πρωτεΐνη), C3 (συμπλήρωματικό συστατικό 3), GNAI2 (G-πρωτεΐνη υπομονάδα άλφα i2), KNG1 (κινινογόνο 1) και SAA1 (αμυλοειδές ορού A1)) είχαν εξαιρετικά υψηλά επίπεδα έκφρασης. Το επίπεδο έκφρασης ενός γονιδίου σε ένα συγκεκριμένο ιστό υπολογίστηκε με τον υπολογισμό του μέσου όρου των λογαριθμικών τιμών του γονιδίου έκφρασης ανά ιστό. Παραδόξως, οι G-πρωτεΐνες εμφανίζουν, κατά μέσο όρο, το υψηλότερο επίπεδο έκφρασης (Εικόνα 6.3). Η φαινομενική αντίφαση, ότι οι τελεστές έχουν μικρότερη έκφραση από τις G-πρωτεΐνες (οι οποίες βρίσκονται πριν στο μονοπάτι μεταγωγής σήματος), μάλλον εξηγείται από τον πολύ μεγαλύτερο αριθμό τελεστών, ενώ οι G-πρωτεΐνες είναι μόλις 13 (η μικρότερη ομάδα στο δίκτυο). Αντιθέτως, οι GPCRs εμφανίζουν το χαμηλότερο επίπεδο έκφρασης σε όλους τους ιστούς (Εικόνα 6.3). Αυτό πιθανώς οφείλεται στο γεγονός ότι διαφορετικοί προσδέτες συνδέονται με διαφορετικούς GPCRs και υπάρχουν και αρκετοί GPCRs οι οποίοι δεν έχουν αλληλεπίδραση με κάποιο προσδέτη (Zhang and Eggert, 2013), οδηγώντας έτσι σε διαφορετικά επίπεδα έκφρασης. Τα υψηλότερα επίπεδα έκφρασης για τις G-πρωτεΐνες και τους τελεστές βρέθηκαν στον υποθάλαμο, ενώ για τους προσδέτες και τους GPCRs βρέθηκαν στο ήπαρ και τον σπλήνα, αντίστοιχα (Εικόνα 6.3). Για τη σχηματική απεικόνιση της γονιδιακής έκφρασης δημιουργήθηκε ένας θερμοχάρτης (Heatmap) στον οποίο συγκρίνεται η έκφραση των 433 γονιδίων ανά ιστό και ανά τύπο μορίου με κλιμακούμενες χρωματικές αλλαγές ανάλογα με τα επίπεδα έκφρασης (Εικόνα 6.4).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

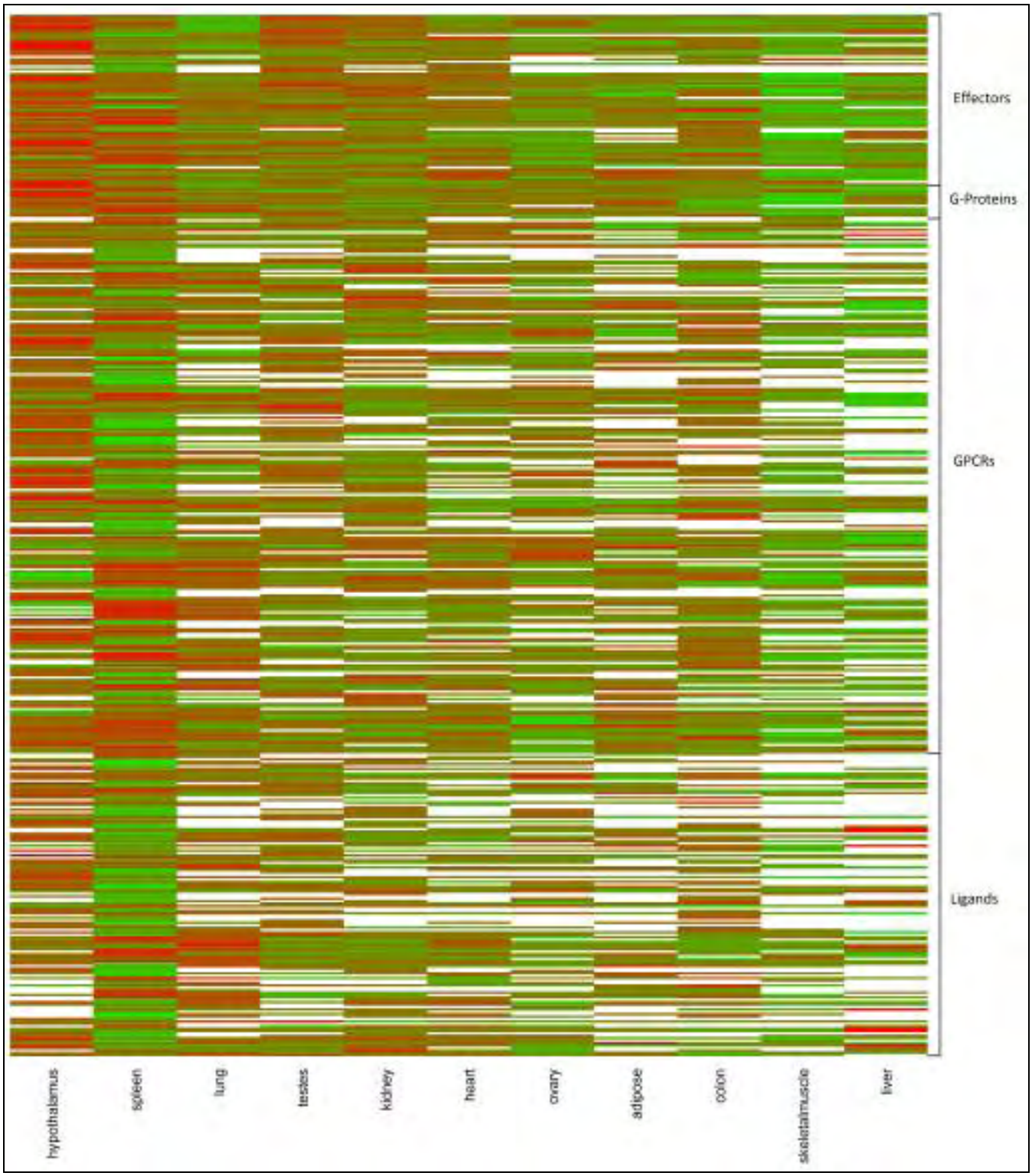


Εικόνα 6.2. Κατανομή των 433 γονιδίων που ανακτήθηκαν από τη βάση δεδομένων Human-gpDB ανάλογα με τον τύπο του μορίου, την παρουσία ή όχι ασθένειας και τον ιστό έκφρασης.



Εικόνα 6.3. Μέσες τιμές του δεκαδικού λογαρίθμου της έκφρασης των γονιδίων ανά ιστό, ανά μόριο και ανά ύπαρξη ή όχι ασθένειας.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

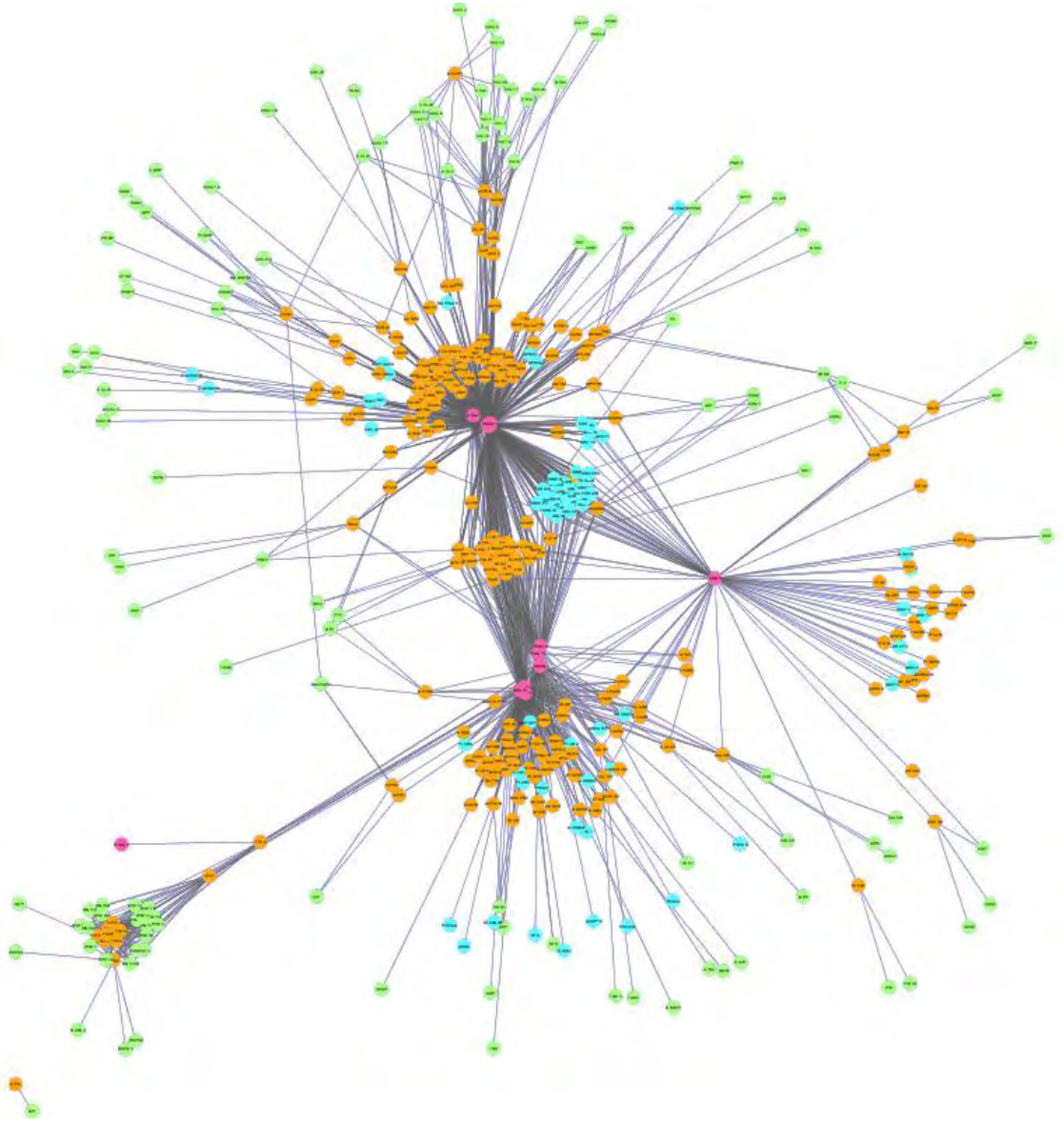


Εικόνα 6.4. Θερμοχάρτης (Heatmap) διαφορικής έκφρασης των 433 γονίδιων ανά ιστό και ανά τύπο μορίου. Το πράσινο δηλώνει μικρή έκφραση, το κόκκινο μεγάλη, ενώ η απουσία έκφρασης δηλώνεται με το λευκό. Στην σπλήνα, εκφράζονται όλα τα γονίδια, αν και ένας μεγάλος αριθμός (133 γονίδια) έχουν ιδιαίτερα χαμηλές εκφράσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

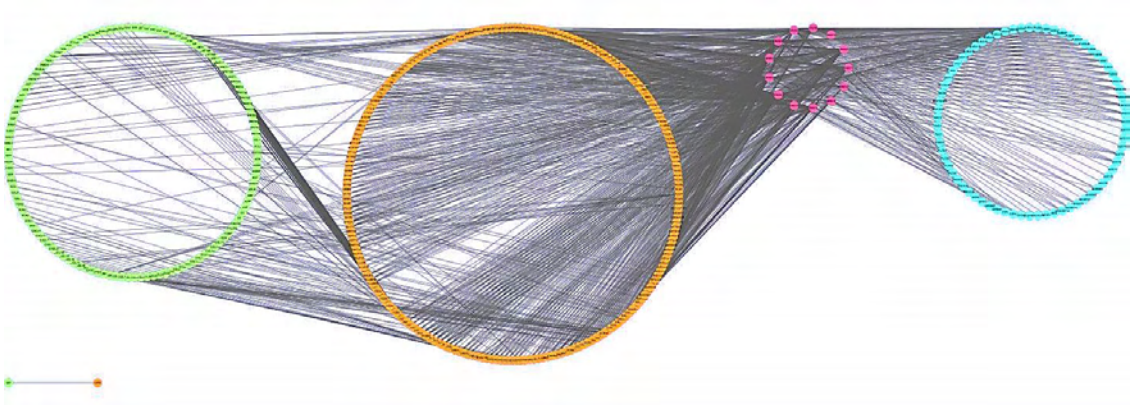
6.3.2 Ανάλυση Δικτύων

Το κυριότερο μέρος της ανάλυσης του συστήματος ligand-GPCR-G-protein-effector εστιάστηκε τόσο στην ανάλυση του κύριου δικτύου όσο και των 11 ειδικών ανά ιστό δίκτυα. Στις παρακάτω εικόνες παρουσιάζεται η τοπολογία του συνολικού δικτύου μεταγωγής σήματος (433 γονίδια, 1666 αλληλεπιδράσεις) με διαφορετικούς αλγορίθμους οπτικοποίησης.

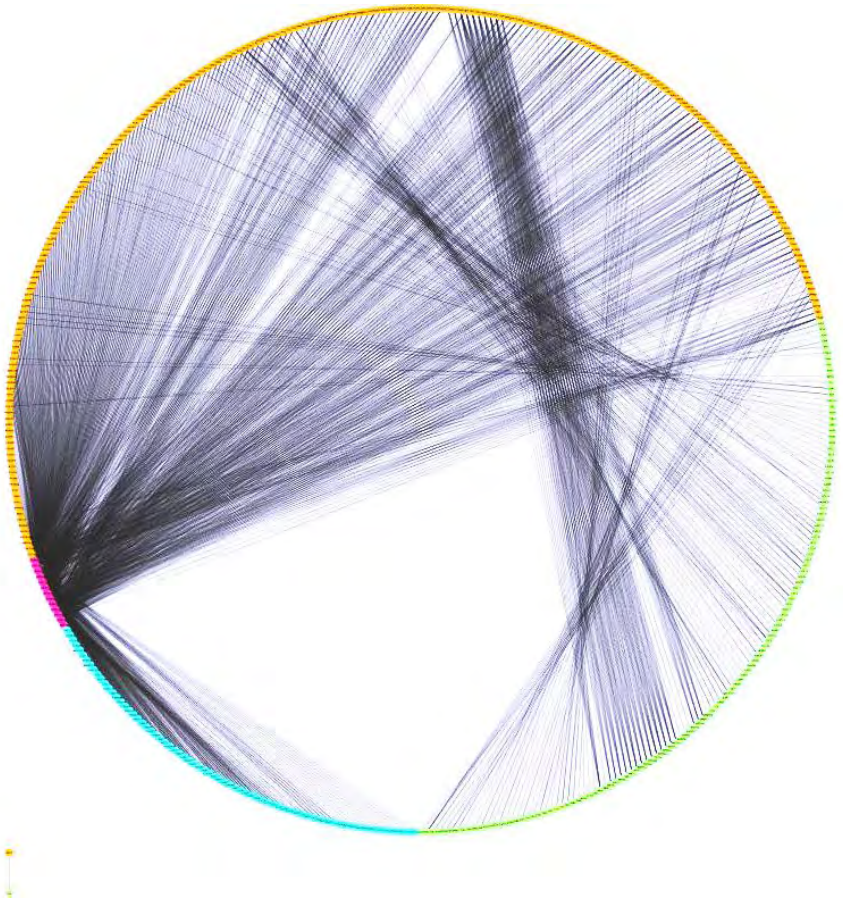


Εικόνα 6.5. Το συνολικό δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεΐνες. Οι κόμβοι είναι χρωματισμένοι ανάλογα με τον τύπο του μορίου. Το πράσινο χρώμα αντιστοιχεί στους συνδέτες (ligands), το πορτοκαλί χρώμα στους GPCRs, το ροζ χρώμα στις G-πρωτεΐνες και το γαλάζιο χρώμα στους τελεστές (effectors).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

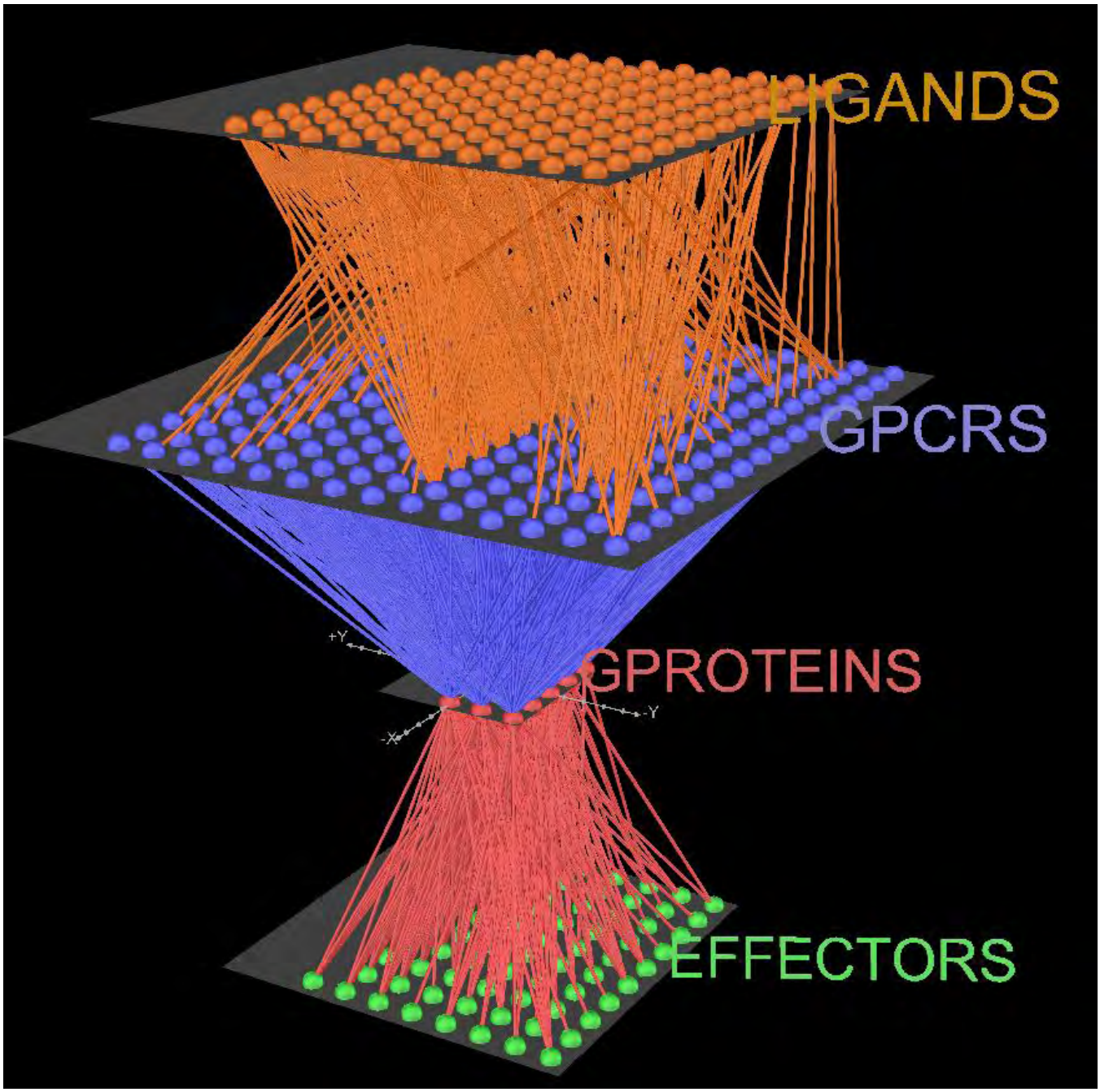


Εικόνα 6.6. Δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεϊνες. Οι κόμβοι είναι χρωματισμένοι ανάλογα με τον τύπο του μορίου. Το πράσινο χρώμα αντιστοιχεί στους ligands, το πορτοκαλί χρώμα στους GPCRs, το ροζ χρώμα στις G-proteins και το γαλάζιο χρώμα στους effectors. Η οπτικοποίση είναι σε διάταξη που απιεκονίζει τη μεταγωγή σήματος.



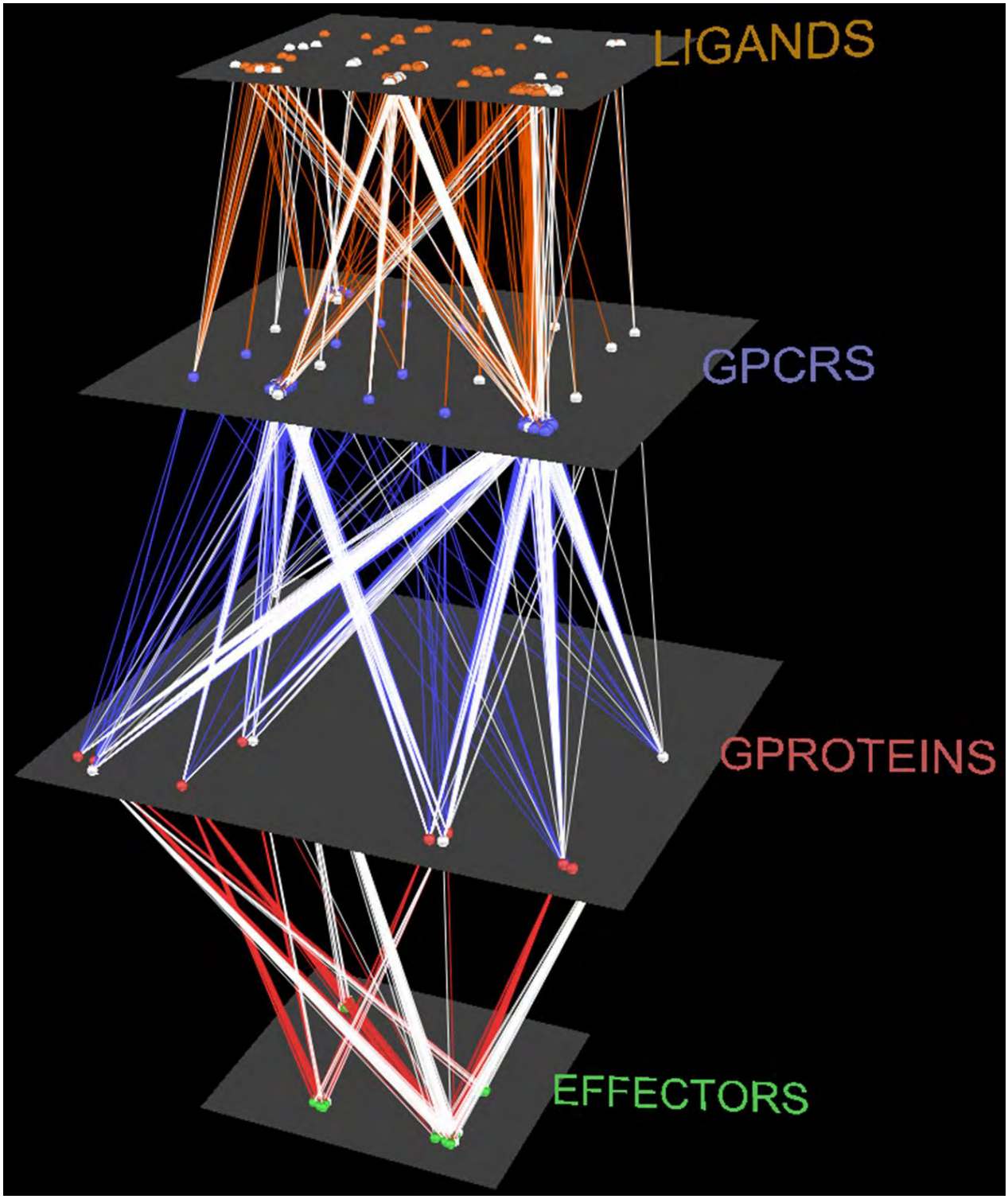
Εικόνα 6.7. Δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεϊνες. Οι κόμβοι είναι χρωματισμένοι ανάλογα με τον τύπο του μορίου. Το πράσινο χρώμα αντιστοιχεί στους συνδέτες, το πορτοκαλί χρώμα στους GPCRs, το ροζ χρώμα στις G-πρωτεΐνες και το γαλάζιο χρώμα στους effectors (κυκλική διάταξη). Εκτός από τις G-πρωτεΐνες για τις οποίες αναμένουμε λόγω μικρού μεγέθους δείγματος οι συνδέσεις να είναι πυκνές, βλέπουμε και άλλες «περιοχές» με αυξημένη πυκνότητα και συνδεσιμότητα (τόσο στους συνδέτες, όσο και στους υποδοχείς και στους τελεστές).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.8. Τρισδιαστατη απεικόνιση του συνολικού δκτύου μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεΐνες. Οι κόμβοι είναι χρωματισμένοι ανάλογα με τον τύπο του μορίου. Το πορτοκαλί χρώμα αντιστοιχεί στους συνδέτες (ligands), το γαλάζιο χρώμα στους GPCRs, το ροζ χρώμα στις G-πρωτεΐνες και το πράσινο χρώμα στους τελεστές (effectors).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.9. Τρισδιαστατη απεικόνιση του συνολικού δκτύου μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεΐνες. Οι κόμβοι είναι χρωματισμένοι ανάλογα με τον τύπο του μορίου. Το πορτοκαλί χρώμα αντιστοιχεί στους συνδέτες (ligands), το γαλάζιο χρώμα στους GPCRs, το ροζ χρώμα στις G-πρωτεΐνες και το πράσινο χρώμα στους τελεστές (effectors). Οι συνδέσεις με λευκό χρώμα απεικονίζουν τη συσχέτιση των γονιδίων με εμπλέκονται σε κάποια ασθένεια.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Το συνολικό δίκτυο είναι ένα κατευθυνόμενο δίκτυο στο οποίο κάθε κόμβος του έχει δύο διαφορετικούς βαθμούς *(k)*, δηλαδή, τον αριθμό των ακμών που συνδέονται με τον κόμβο. Ο εξερχόμενος βαθμός ενός κόμβου (k_{out}) δείχνει τον αριθμό των εξερχόμενων ακμών που προέρχονται από έναν κόμβο, και ο εισερχόμενος βαθμός (k_{in}) είναι ο αριθμός των εισερχόμενων ακμών σε έναν κόμβο. Ο συνολικός βαθμός του κόμβου (k_{tot}) είναι το άθροισμα των επιμέρους βαθμών ($k_{out} + k_{in}$) (Barabasi and Oltvai, 2004). Η κατανομή του βαθμού των κόμβων $P(k)$ του δικτύου που έχει k συνδέσεις με άλλους κόμβους ορίζεται από τη σχέση $P(k) \sim k^{-\gamma}$, όπου γ η εκτιμώμενη παράμετρος του νόμου δύναμης για την κατανομή του βαθμού των κόμβων (Barabasi, 2009; Barabasi and Albert, 1999). Το δίκτυο είναι ένα κατευθυνόμενο δίκτυο και ανήκει στην κατηγορία των δικτύων άνευ κλίμακας καθώς ακολουθεί ξεκάθαρα την κατανομή νόμου δύναμης (Barabasi, 2009). Στο δίκτυο υπάρχουν λίγοι κόμβοι οι οποίοι εμφανίζουν μεγάλο αριθμό συνδέσεων (κεντρικοί κόμβοι). Στην Εικόνα 6.10 παρουσιάζεται η κατανομή του βαθμού των κόμβων η οποία επιβεβαιώνει ότι ακολουθεί την κατανομή νόμου δύναμης. Επίσης, όσο πιο μικρή είναι η τιμή του εκθέτη του βαθμού (γ) τόσο περισσότερο οι ιδιότητες του δικτύου καθορίζονται από τους κεντρικούς κόμβους. Αυτό εξηγείται καπό τις τιμές γ που παρουσιάζονται στον Πίνακα 6.1.

Κατά παρόμοιο τρόπο, τα ειδικά για κάθε ιστό δίκτυα παρήχθησαν λαμβάνοντας υπόψη το προφίλ έκφρασης των γονιδίων που κωδικοποιούν τους τέσσερις τύπους μορίων σε κάθε ιστό. Τα έντεκα ειδικά ανά ιστό δίκτυα εμφανίζουν την ίδια αρχιτεκτονική (Barabasi, 2009) με το συνολικό δίκτυο (Εικόνες 6.10-6.11) υποδεικνύοντας ότι ο τρόπος δράσης και η μεταγωγή σήματος από τους GPCR, γίνεται με περίπου σταθερό τρόπο σε κάθε ιστό. Ένας αλγόριθμος μη επιβλεπόμενης ομαδοποίησης (Enright et al., 2002) χρησιμοποιήθηκε για την ανίχνευση των λειτουργικών μονάδων σε αυτά τα δίκτυα. Σε όλους τους ιστούς φαίνεται να σχηματίζεται μια μεγάλη κυρίαρχη ομάδα με τα περισσότερα γονίδια ενώ τα υπόλοιπα ταξινομούνται σε ομάδες με μικρό αριθμό γονιδίων. Αυτό είναι κάτι αναμενόμενο, καθώς τα δίκτυα άνευ κλίμακας, δεν μπορούν εύκολα να σχηματίσουν ομάδες (Barabasi, 2009). Τα 11 δίκτυα είναι ελαφρώς πυκνότερα σε σύγκριση με το κοινό δίκτυο, δεδομένου ότι οι τιμές της ολικής σύνδεσης/πυκνότητας (connectivity/density) και κεντροποίησης (centralization) (Πίνακας 6.1) είναι οριακά υψηλότερες σε σύγκριση με αυτές του συνολικού δικτύου (Barabasi and Oltvai, 2004). Αυτό σημαίνει ότι περισσότερες ακμές συνδέονται με έναν κόμβο σε αυτά τα δίκτυα σε σύγκριση με τον ίδιο κόμβο στο συνολικό δίκτυο. Οι παραπάνω παρατηρήσεις οδηγούν στην υπόθεση ότι τα συγκεκριμένα γονίδια υπερεκφράζονται σε συγκεκριμένους ιστούς και συνδέονται μεταξύ τους για να σχηματίσουν ένα σύμπλοκο που εμπλέκεται στις λειτουργίες τους.

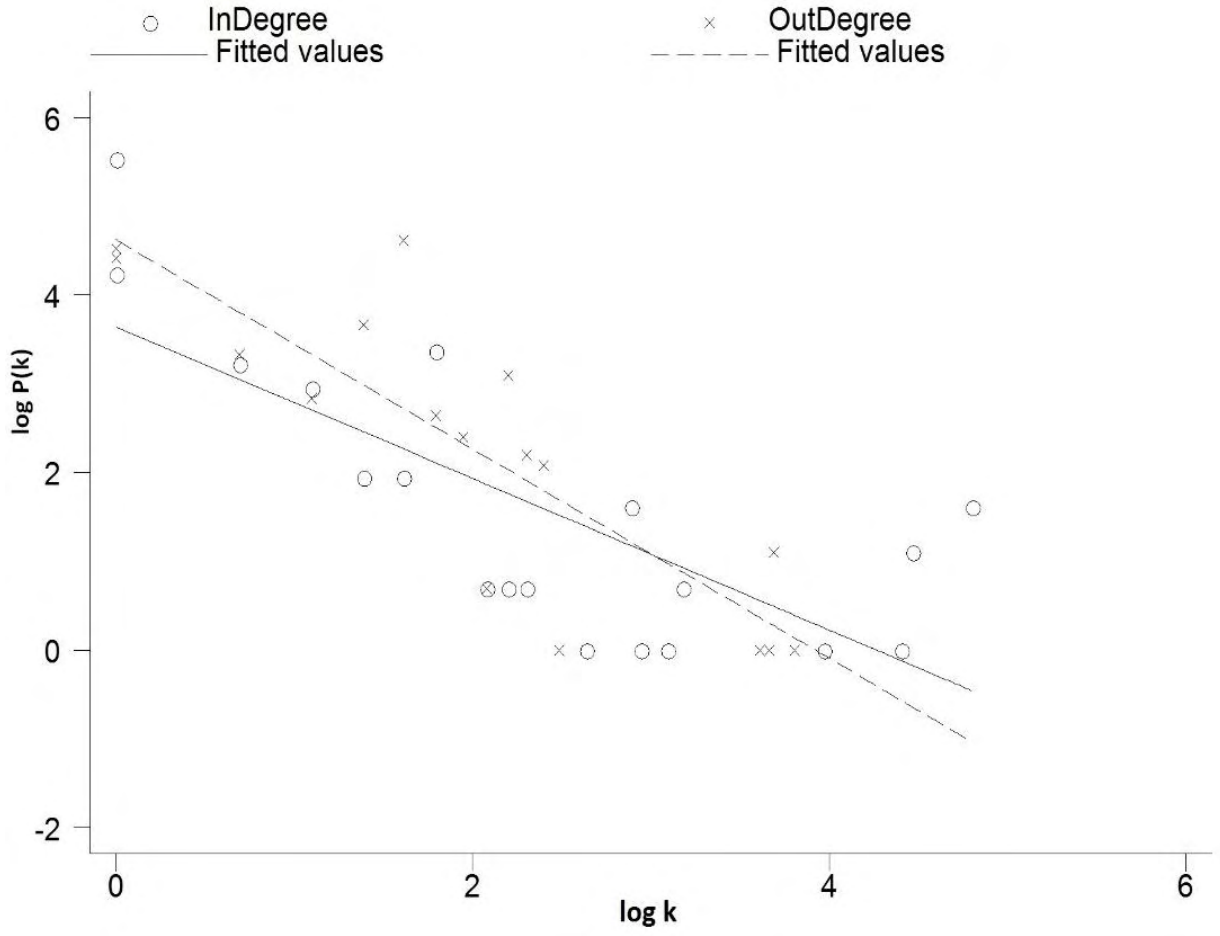
Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Όλοι οι κόμβοι του συνολικού δικτύου ταξινομήθηκαν με βάση το συνολικό βαθμό συνδεσιμότητας. Οι δέκα κόμβοι με την υψηλότερη συνδεσιμότητα παρουσιάζονται στον Πίνακα 6.2. Μεταξύ αυτών, το 40% των συνδετών και G-πρωτεϊνών, καθώς επίσης και το 30% των GPCRs και τελεστών, συσχετίζονται με ασθένειες. Επιπλέον, οι G-πρωτεΐνες GNAI2, GNAI3 και GNAS είχαν, κατά πολύ, την υψηλότερη συνολική κατανομή βαθμού *(ktot)* σε όλους τους ιστούς (Πίνακας 6.2).

Πίνακας 6.1. Στατιστικά μέτρα των δικτύων**.**

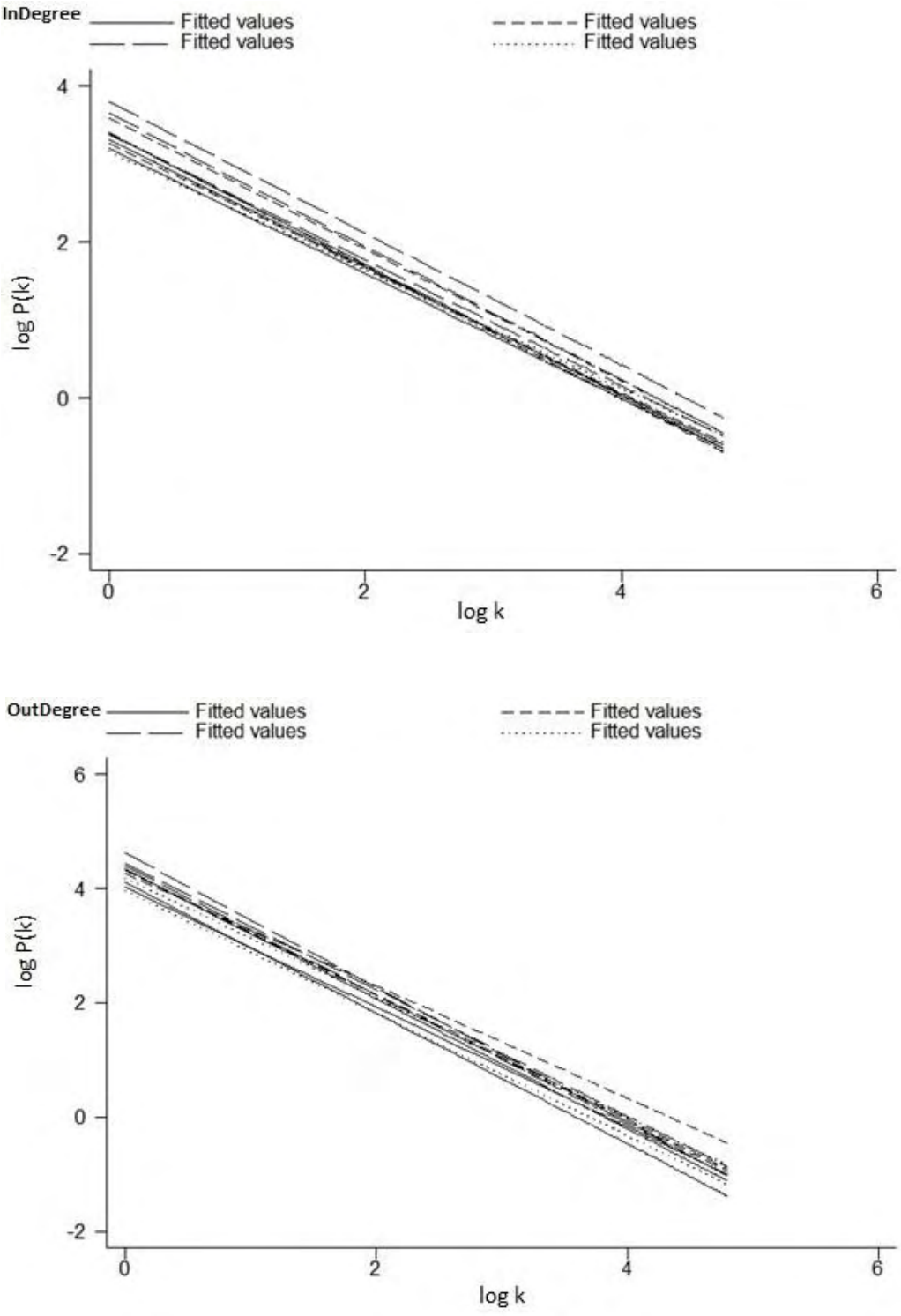
	Κόμβοι	**Ακμές**	**Μήκος Μονοπατιού**	**γ (εισ)**	**γ (εξ)**	**Συνολική Συνδεσιμότητα / Πυκνότητα**	**Συντελεστής Κεντροποίησης**	**Μέγιστος αριθμός ακμών**
Λιπώδης Ιστός	292	1073	2.09	0.70	1.01	0.013	0.29	86
Παχύ έντερο	330	1199	2.10	0.71	1.09	0.011	0.29	96
Καρδιά	328	1233	2.13	0.59	1.02	0.012	0.29	97
Υποθάλαμος	363	1357	2.15	0.71	1.13	0.010	0.28	105
Νεφρό	356	1270	2.15	0.65	1.11	0.010	0.28	99
Συκώτι	250	885	2.08	0.70	1.03	0.015	0.28	72
Πνεύμονας	337	1254	2.15	0.66	1.14	0.011	0.28	97
Ωοθήκη	325	1142	2.13	0.65	1.10	0.011	0.30	99
Σκελετικός Μυς	279	923	2.08	0.65	1.13	0.013	0.31	84
Σπλήνα	323	1228	2.12	0.57	1.00	0.010	0.28	97
Όρχεις	350	1338	2.16	0.68	0.97	0.011	0.30	108
Συνολικό Δίκτυο	**433**	**1666**	**2.22**	**0.70**	**1.20**	**0.008**	**0.27**	**122**

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.10. Κατανομή του αριθμού των κόμβων για το συνολικό δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεϊνες.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.11. Κατανομή του αριθμού των κόμβων για τα δίκτυα μεταγωγής σήματος των ανθρώπινων υποδοχέων συζυγμένων με G-πρωτεϊνες από καθένα από τους 11 ιστούς.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

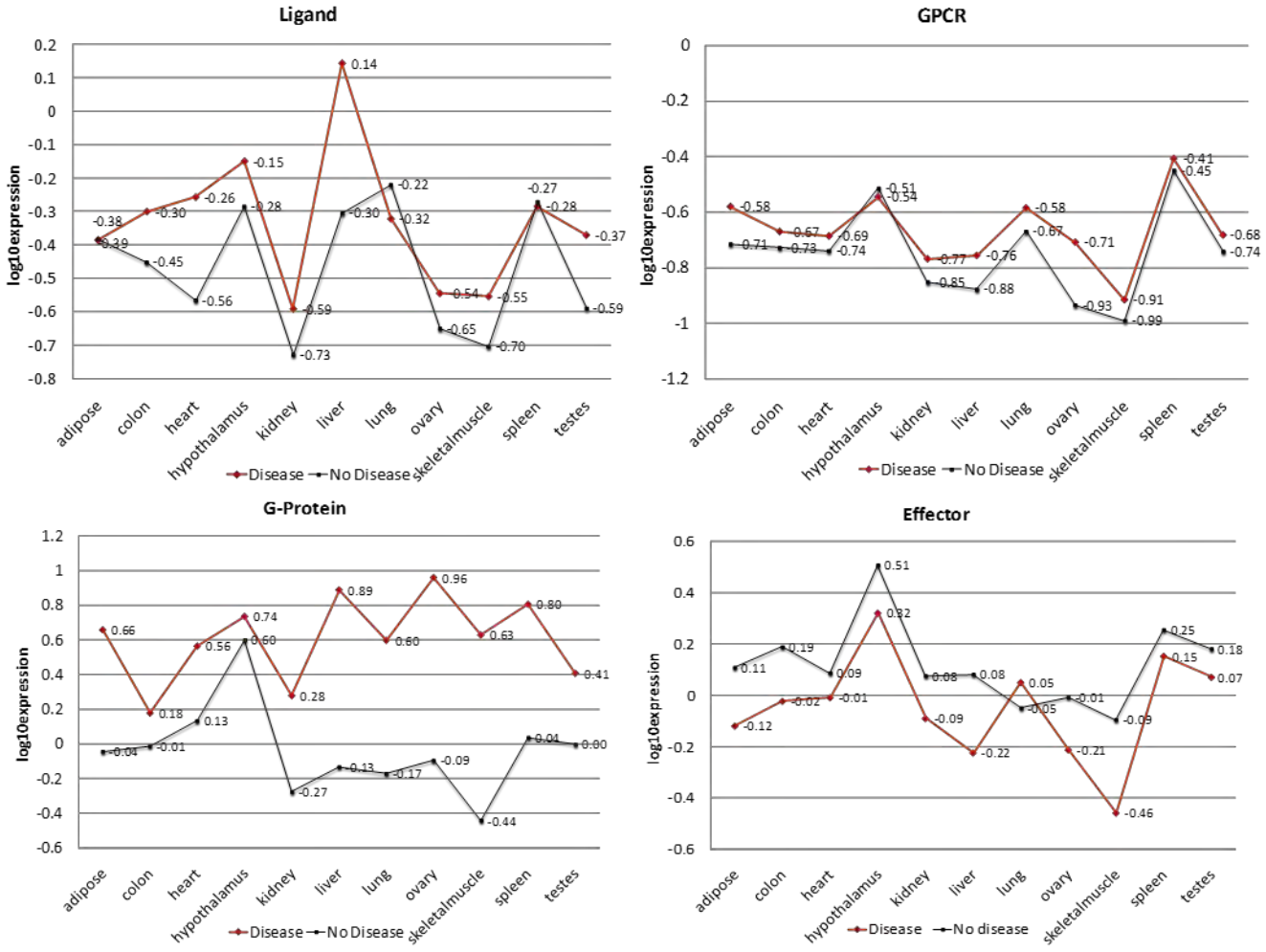
Πίνακας 6.2. Οι 10 κόμβοι με το μεγαλύτερο αριθμό συνδέσεων ανάλογα με τον τύπο του μορίου. Με αστερίσκο (*) συμβολίζονται τα γονίδια που συσχετίζονται με κάποια ασθένεια. Παρατηρούμε ότι το FZD3 περιγράφεται ως μή έχον σχέση με ασθένεια, καθώς η μελέτη μας που περιγράφηκε στο Κεφάλαιο 2, δεν είχε δημοσιευθεί και δεν περιλαμβανόταν στην GAD. Αυτό το γεγονός επιβεβαιώνει την ανάγκη για μια πιο σύγχρονη και συνεχώς ανανεούμενη βάση δεδομένων για τις γενετικές συσχετίσεις.

Ligands	**GPCRS**	**G-proteins**	**Effectors**
WNT11*	FZD1	GNAI1	TUBA1A
WNT2B	FZD8	GNAI2*	TUBA1B
WNT4*	CCR10	GNAI3*	TUBA1C
WNT5A	FZD4*	GNAO1	TUBA3C
WNT5B	FZD3	GNAQ	TUBA3D
WNT16*	FZD5	GNAZ	TUBA3E
WNT3*	FZD6	GNA11	TUBA4A
WNT7B	FZD7	GNAS*	TUBB3*
WNT9A	TSHR*	GNA14	TUBA8*
WNT9B	F2R*	GNA15*	TUBB1*

6.3.3 Προφίλ έκφρασης γονιδίων που συσχετίζονται με ασθένειες

Όπως αναφέρθηκε παραπάνω, τα δεδομένα για τις συσχετίσεις γονίδιου-ασθένειας ανακτήθηκαν από το δίκτυο γονιδίων ασθένειων που δημιουργήθηκε στο προηγούμενο Κεφάλαιο (Kontou et al., 2016d). Περίπου το ένα τρίτο (28,6%) του συνολικού αριθμού των γονιδίων που κωδικοποιούν τους τέσσερις τύπους μορίων βρέθηκε να σχετίζεται με μια ή περισσότερες ασθένειες (Εικόνα 6.2). Ασθένειες του ενδοκρινολογικού συστήματος, μεταβολικές ασθένειες και ασθένειες του κυκλοφορικού συστήματος ήταν οι πιο συχνές. Σε επόμενο βήμα προσπαθήσαμε να διερευνήσουμε αν η υποέκφραση ή η υπερέκφραση ορισμένων τύπων μορίων σε συγκεκριμένους ιστούς σε φυσιολογικές συνθήκες (υγιή άτομα) μπορεί να συνδέεται με αυξημένη πιθανότητα εμφάνισης των ασθενειών στις οποίες να εμπλέκονται αυτά τα γονίδια. Σε κάθε τύπο μορίου βρέθηκαν ότι εμφανίζονται μεγαλύτερα επίπεδα έκφρασης των γονιδίων σε παθολογικές καταστάσεις με εξαίρεση τους τελεστές για τους οποίους φαίνεται να ισχύει αντίστροφη σχέση (Εικόνα 6.12).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.12. Μέση τιμή του δεκαδικού λογαρίθμου της έκφρασης των τεσσάρων κατηγοριών μορίων ανά ιστό και ανάλογα με την εμφάνιση ασθένειας.

6.3.4 Φάρμακα και αλληλεπιδράσεις των μορίων του δικτύου των GPCR

Για κάθε τύπο μορίου (προσδέτες, GPCRs, G-πρωτεΐνες και τελεστές), ανακτήθηκε ο αριθμός των φαρμάκων που τους στοχεύουν και ο συνολικός αριθμός των αλληλεπιδράσεων με άλλα μόρια. Ο μεγαλύτερος αριθμός των φαρμάκων καταγράφηκε για τους GPCRs, με μέση τιμή περίπου ίση με 8, με ένα εύρος τιμών να κυμαίνεται από 0 έως 85. Κανένα φάρμακο δε βρέθηκε να στοχεύει τις G-πρωτεΐνες. Ο αριθμός των φαρμάκων που στοχεύουν τους προσδέτες και τους τελεστές κυμαίνεται από 0 έως 5. Επιπλέον, οι τελεστές βρέθηκαν να έχουν τον μεγαλύτερο αριθμό αλληλεπιδράσεων πρωτεϊνών-πρωτεϊνών (PPIs) με μια μέση τιμή περίπου ίση με 28. Για τις G-πρωτεΐνες η μέση τιμή του αριθμού των αλληλεπιδράσεων ήταν ίση με 12.5, για τους προσδέτες η μέση τιμή ήταν ίση με 6 ενώ ο αριθμός των αλληλεπιδράσεων κυμαίνονταν από 0 έως 77 για τους GPCRs.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

6.3.5 Μοριακά Μονοπάτια

Προκειμένου να βρεθεί ο τρόπος με τον οποίο εμπλέκονται τα γονίδια του δικτύου μεταγωγής σήματος σε ασθένειες, μελετήθηκε το συνολικό δίκτυο και βρέθηκαν σε αυτό μονοπάτια στα οποία κάνενας από τους 4 κόμβούς τους δεν εμπλέκεται σε κάποια ασθένεια (ένα χαρακτηριστικό παράδειγμα φαίνεται στην Εικόνα 6.13B), βρέθηκαν μονοπατια που 1 ή 2 κόμβοι τους συσχετίζονται με κάποια ασθένεια και τέλος, βρέθηκαν και λίγα μονοπάτια στα οποία και οι 4 κόμβοι τους συσχετίζονται με διάφορες ασθένειες. Χαρακτηριστικό είναι το ακόλουθο μονοπάτι στο οποίο οι κόμβοι του εμπλέκονται σε ασθένεια της ίδιας κατηγορίας (Εικόνα 6.13A).

Τα γονίδια που κωδικοποιούν το μοριακό μονοπάτι της Εικόνα 6.13B εκφράζονται στην καρδιά, γεγονός που υποδηλώνει ότι τα αυξημένο επίπεδα έκφρασης αυτών των γονιδίων είναι απαραίτητα για τη φυσιολογική δραστηριότητα της καρδιάς. Αυτά τα γονίδια βρέθηκαν να ρυθμίζουν ποικίλες καρδιαγγειακές λειτουργίες. Για παράδειγμα, το γονίδιο CXCL11 (χημειοκίνη) έχει βρεθεί να παίζει σηματικό ρόλο σε μοντέλα μεταμόσχευσης καρδιάς (Mitsuhashi et al., 2007) και το γονίδιο ACKR3 (Άτυπος χημειοκινικός υποδοχέας 3) εμπλέκεται στην αγγειοσυστολή (Bach et al., 2014). Η G-πρωτεΐνη GNAI1 εμπλέκεται στη ρύθμιση της δυναμικής του καρδιακού ρυθμού (Zuberi et al., 2008). Ο τελεστής TUBGCP3 είναι απαραίτητος για τη λειτουργία των μικροσωληνίσκων (Oakley, 1992). Οι μικροσωληνίσκοι, μαζί με την ακτίνη, συνθέτουν το κύριο συστατικό του κυτταροσκελετού των καρδιομυοκυττάρων (Sequeira et al., 2014). Δεδομένου ότι ο τελεστής TUBGCP3 είναι απαραίτητος στην καρδιά, προτείνεται ότι μπορεί να διαδραματίσει ένα σημαντικό ρόλο στον καρδιακό κυτταροσκελετό.

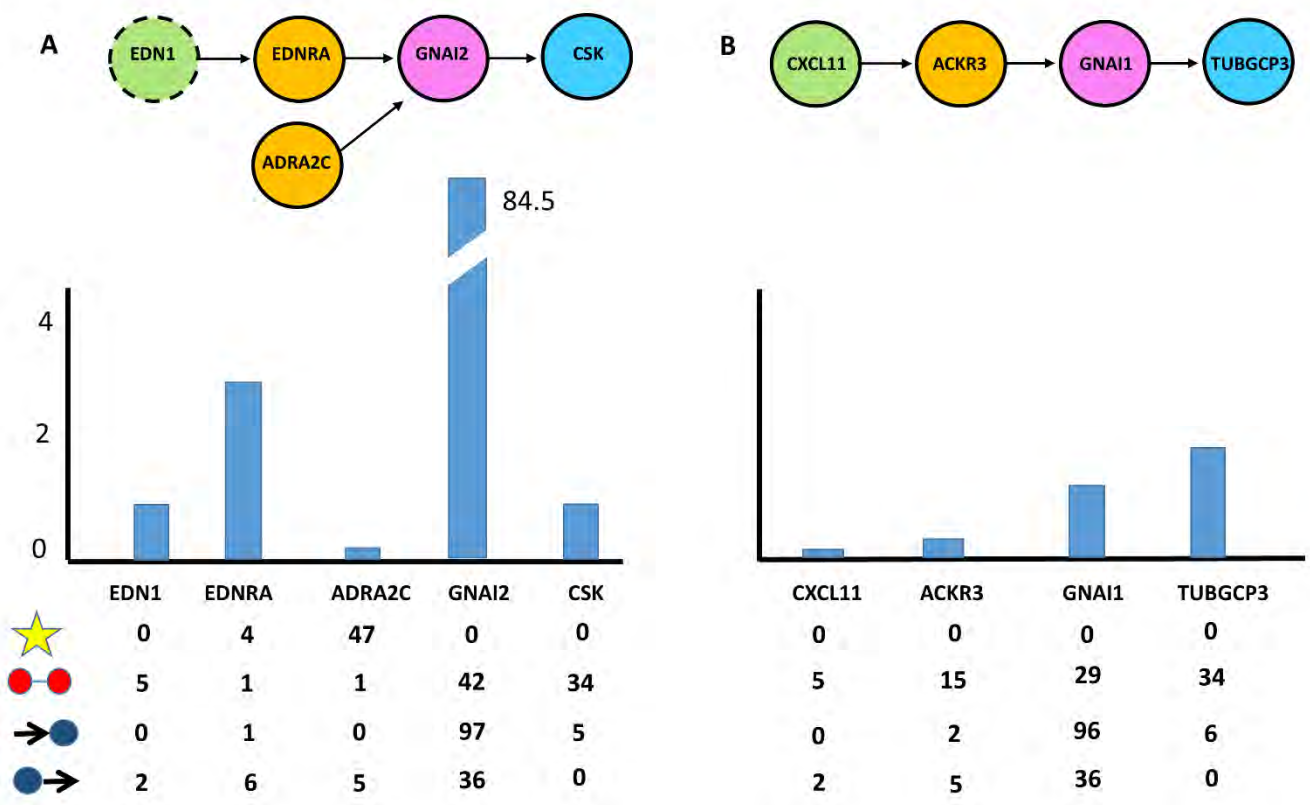
Τα μόρια που συμμετέχουν στο μονοπάτι που φαίνεται στην Εικόνα 6.13A είναι εντόνως εκφρασμένα στον καρδιακό ιστό κάτι το οποίο δείχνει ότι τα συγκεκριμένα γονίδια μπορεί να προκαλέσουν κάποια καρδιακή παθολογία, όπως οι καρδιαγγειακές παθήσεις. Το γονίδιο *EDN1* (Endothelin-1), που κωδικοποιεί τον προσδέτη ενδοθηλίνη-1, πιστεύεται ότι συμβάλλει στην παθογένεση της υπέρτασης, της καρδιακής ανεπάρκειας και της αθηροσκλήρωσης (Lifton et al., 2001). Επιπλέον, τα επίπεδα λιποπρωτεΐνης υψηλής πυκνότητας (HDL, High-Density Lipoprotein) ή HDL χοληστερόλης συσχετίζονται με μια αντικατάσταση K198N στο γονίδιο *EDN1* (Pare et al., 2007). Πολυμορφισμοί στο γονίδιο *EDNRA* (endothelin receptor type A) που κωδικοποιεί τον υποδοχέα ενδοθηλίνης τύπου Α σχετίζονται με πνευμονική αρτηριακή υπέρταση. Ο υποδοχέας EDNRA, που δεν ενεργοποιείται από κάποιον ειδικό προσδέτη, μεταδιεγείρει το μεγαλύτερο μέρος των αγγειοσυσπαστικών ιδιοτήτων του προσδέτη ενδοθηλίνης-1 (Benjafield et al., 2003). Το γονίδιο

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ADRA2C (Adrenoceptor Alpha 2C) κωδικοποιεί τον Αδρενο-υποδοχέα Άλφα 2C, ο οποίος απαιτείται για τη ρυθμιζόμενη απελευθέρωση των νευροδιαβιβαστών από νοραδρενεργικούς νευρώνες στην καρδιά. Πολυμορφισμοί στο γονίδιο *ADRA2C* σχετίζονται με την υπέρταση (Savva et al., 2012). Η αναστολή της G-πρωτεΐνης, που δεσμεύει νουκλεοτίδια της γουανίνης, κωδικοποιείται από το γονίδιο *GNAI* (Guanine nucleotide binding protein 2) και ενεργοποιείται και από τους δύο υποδοχείς, EDNRA και ADRA2C. Προτείνεται ότι η GNAI εμπλέκεται στην παθογένεση της ιδιοπαθούς υπέρτασης. Έχει δειχθεί ότι άτομα με μονονουκλεοτιδικούς πολυμορφισμούς στο γονίδιο *GNAI2* έχουν αυξημένο κίνδυνο να εμφανίσουν υπέρταση (Menzaghi et al., 2006). Πολυμορφισμοί στο γονίδιο που κωδικοποιεί τον τελεστή *CSK* (C-Src tyrosine kinase), C-Src κινάση της τυροσίνης, συσχετίζονται σε σημαντικό βαθμό με αυξημένο κίνδυνο υπέρτασης (Xi et al., 2013). Παρόλο που ο προσδέτης εμπλέκεται σε διαφορετικό μονοπάτι από αυτό των GPCRs, της G-πρωτεΐνης και του τελεστή, τα δύο μονοπάτια συνδέονται με σημαντικό τρόπο, καθώς έχει δειχθεί σε αρκετές μελέτες ότι η HDL χοληστερόλη συσχετίζεται θετικά με την υπέρταση (Oda and Kawai, 2011).

Τα δεδομένα αυτά, είναι ιδιαίτερα σημαντικά, καθώς με την ανάλυση μας, δείχνουμε ότι τα γονίδια αυτά είναι πιθανό να εμπλέκονται και σε άλλους φαινότυπους όπως για παράδειγμα σε γενικότερους φαινότυπους του μεταβολικού συνδρόμου (Alberti et al., 2005), όπως LDL χοληστερίνη, υπερλιπιδαιμία και διαβήτη τύπου ΙΙ, και εισηγούνται στοχευμένες μελέτες εύρεσης της αλληλεπίδρασης γονιδίου-γονιδίου σε μελέτες γενετικής συσχέτισης (gene-gene interaction) (Manning et al., 2011). Το μεγάλο πρόβλημα με τις μελέτες αυτές, είναι ότι σε γονιδιωματική κλίμακα, η ταυτόχρονη ανάλυση όλων των ζευγαριών γονιδίων είναι και δύσκολη υπολογιστικά αλλά και επίφοβη στατιστικά, καθώς το σφάλμα τύπου Ι, θα είναι υπερβολικά μεγάλο λόγω των πολλών ελέγχων (multiple comparison).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.13. (Α) Παράδειγμα μονοπατιών στα οποία όλα τα τέσσερα μόρια που εμπλέκονται στην ίδια κατηγορία νόσου. (Β) Παράδειγμα μονοπατιών όπου τα μόρια δεν εμπλέκονται σε ασθένειες. Συνδέτης (πράσινο χρώμα), GPCRs (πορτοκαλί χρώμα), G-πρωτεΐνη (ροζ χρώμα) και τελεστές (μπλε χρώμα). Το δεύτερο μονοπάτι δεν περιέχει συνδέτη. Τα ιστογράμματα δείχνουν την έκφραση κάθε γονιδίου που συμμετέχουν σε κάθε μονοπάτι. Με κίτρινο αστερίσκο απεικονίζεται ο αριθμός των φαρμάκων που αλληλεπιδρούν με κάθε γονίδιο/πρωτεΐνη, με συνδεδεμένες τελείες (κόκκινο χρώμα) απεικονίζεται ο συνολικός αριθμός των PPIs στην οποία εμπλέκεται η αντίστοιχη πρωτεΐνη, με βέλος-κουκκίδα (μπλε) απεικονίζεται ο αριθμός γονιδίων που ενεργοποιούν το αντίστοιχο γονίδιο στο δίκτυο (εισερχόμενες αλληλεπιδράσεις), με κουκκίδα-βέλος (μπλε) απεικονίζεται ο αριθμός γονιδίων που ενεργοποιούνται από το αντίστοιχο γονίδιο (εξερχόμενες αλληλεπιδράσεις).

6.3.6 Πρότυπα συν-έκφρασης γονιδίων (Gene co-expression patterns)

Τα γονίδια των αλληλεπιδρώντων μορίων που εκφράζονται στον ίδιο ιστό ονομάζονται συν-εκφρασμένα γονίδια (Heyer et al., 1999) και στο εξής για την καλύτερη κατανόηση θα αναφέρονται ως ζεύγη γονιδίων. Τα ζεύγη γονιδίων στο συνολικό δίκτυο είναι: προσδέτης-GPCR (20,6%), GPCR-G-πρωτεΐνη (63,3%) και G-πρωτεΐνη-τελεστής (16,1%). Τα πρότυπα γονιδιακής συν-έκφρασης διερευνήθηκαν τόσο σε μη παθολογικές όσο και σε παθολογικές καταστάσεις. Περισσότερα από τα μισά ζεύγη γονιδίων (55%) είχαν συσχέτιση με μία ασθένεια (δηλαδή είτε το ένα ή και τα δύο γονίδια

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

του ζεύγους γονιδίων συσχετίζεται με μια ασθένεια). Σε 1 από τα 10 ζεύγη γονιδίων, και τα δύο γονίδια συσχετίζονται με μια ασθένεια. Μεταξύ αυτών, μόνο τρία ζεύγη γονιδίων βρέθηκαν να συσχετίζονται με την ίδια ασθένεια/κατηγορία ασθένειας: το ζεύγος γονιδίων AGT-AGTR1 (προσδέτης-υποδοχέας), σχετίζεται με διαταραχές που προκύπτουν από κακή νεφρική λειτουργία, το ζεύγος γονιδίων GRM7 - GNAI3 (υποδοχέας-G-πρωτεΐνη) σχετίζεται με την κατάθλιψη, το ζεύγος LPAR1-GNA12 (υποδοχέας-G-πρωτεΐνη) συνδέεται με ανωμαλίες στο ύψος. Το επίπεδο έκφρασης των ζευγών γονιδίων υπολογίστηκε από το γινόμενο των τιμών έκφρασης των αλληλεπιδρώντων γονιδίων. Αποδείχθηκε ότι αυτές οι τιμές ποικίλουν ανάλογα με την παρουσία ή απουσία ασθένειας, τον τύπο των αλληλεπιδρώντων μορίων (π.χ. προσδέτης-GPCR, κ.λπ.) και το είδος του ιστού. Με άλλα λόγια, υπάρχει μια αλληλεπίδραση τριών δρόμων (three-way interaction). Ειδικότερα, στο σκελετικό μυϊκό ιστό βρέθηκαν τα χαμηλότερα επίπεδα έκφρασης και στον υποθάλαμο βρέθηκαν τα υψηλότερα επίπεδα έκφρασης. Όπως ήταν αναμενόμενο, τα ζεύγη G-πρωτεΐνες-τελεστές παρουσιάζουν τα υψηλότερα επίπεδα έκφρασης σε σύγκριση με τα άλλα ζεύγη γονιδίων. Συνολικά, τα επίπεδα έκφρασης των ζευγών γονιδίων (GPCR-G πρωτεΐνη και G πρωτεΐνη-τελεστής) που συνδέονται με ασθένειες εμφανίζονται να είναι σταθερά υψηλότερα σε σύγκριση με εκείνα των αντίστοιχων ζευγών γονιδίων που δεν εμπλέκονται σε ασθένειες σε όλους τους ιστούς. Επιπλέον, υπάρχει μια θετική συσχέτιση μεταξύ των προτύπων γονιδίων συν-έκφρασης και των κλινικών εκδηλώσεων των ασθενειών. Για παράδειγμα, το ζεύγος γονιδίων AGT - AGTR1 εκφράζεται στο νεφρό συσχετίζεται με μια νεφρική νόσο.

6.3.7 Μοντέλο εκτίμησης κινδύνου εμφάνισης ασθένειας (Disease risk prediction model)

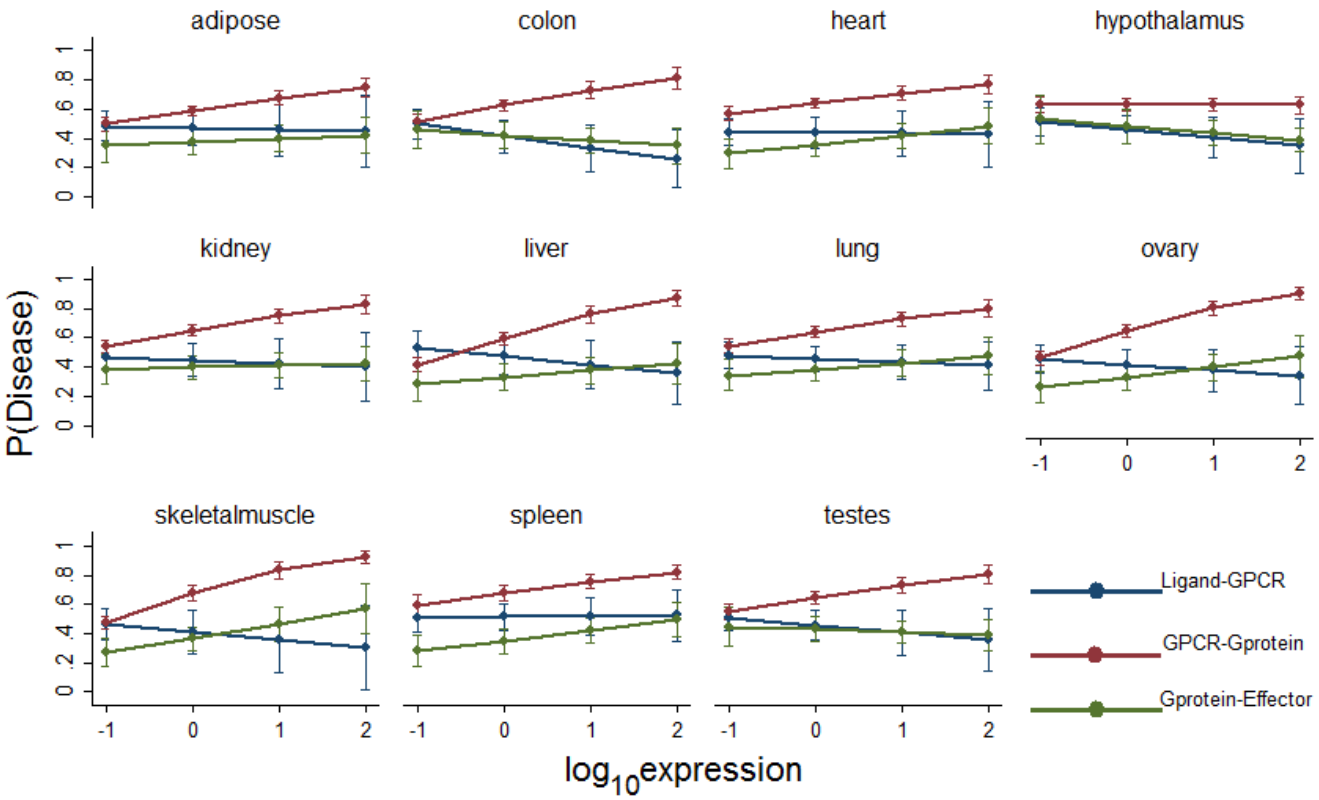
Ένα μοντέλο λογιστικής παλινδρόμησης κατασκευάστηκε για να προβλέψει το σχετικό κίνδυνο ανάπτυξης μιας ασθένειας στην περίπτωση των ζευγαριών γονιδίων (συν-εκφρασμένων γονίδιων). Τα ζευγάρια των γονιδίων επελέγησαν για την ανάλυση, αντί των απλών γονιδίων, προκειμένου να αποτυπωθεί η ροή και η κατεύθυνση της πληροφορίας από το ένα επίπεδο στο άλλο (δηλαδή από τους προσδέτες προς τους GPCRs και ούτω καθεξής). Για το σκοπό αυτό, υπολογίστηκαν τα τυπικά σφάλματα, έτσι ώστε να ληφθούν υπόψη οι εξαρτήσεις που προκύπτουν από το γεγονός ότι προβλέπονται οι τιμές έκφρασης από 11 υγιείς ιστούς εντός του ίδιου ζεύγους γονιδίων. Στο μοντέλο επιπλέον αξιοποιήθηκε η πληροφορία για τον τύπο των αλληλεπιδρώντων μορίων, τον ιστό, τη λογαριθμική μέση τιμή του γινομένου των τιμών έκφρασης των αλληλεπιδρώντων γονιδίων, το άθροισμα του αριθμού των φαρμάκων, το άθροισμα του συνολικού αριθμού των αλληλεπιδράσεων με

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

άλλες πρωτεΐνες (PPIs) και το άθροισμα των εξερχόμενων ακμών που προέρχονται από ένα κόμβο και των εισερχόμενων ακμών πάνω σε ένα κόμβο ανά ιστό των αλληλεπιδρώντων μορίων.

Μέσω του μοντέλου λογιστικής παλινδρόμησης εκτιμήθηκε ότι οι αλληλεπιδράσεις μεταξύ των μορίων GPCRs-G πρωτεΐνες και G πρωτεΐνες-τελεστές εμφανίζουν τη μεγαλύτερη πιθανότητα εμφάνισης ασθενειών η οποία μάλιστα μεγαλώνει όσο αυξάνονται τα επίπεδα έκφρασης των γονιδίων, φτάνοντας στο 0.9 για την πιο ακραία τιμή έκφρασης (Εικόνα 6.14). Ο τύπος του ιστού στον οποίο εκφράζεται το ζεύγος των γονιδίων, δε φαίνεται να τροποποιεί τον κίνδυνο εμφάνισης νόσου με κάπως χαμηλότερες πιθανότητες να εκτιμώνται για τον υποθάλαμο. Ενδιαφέρον παρουσιάζει μια αρνητική συσχέτιση που εμφανίστηκε μεταξύ της πιθανότητας των αλληλεπιδρώντων μορίων προσδέτης-GPCR που εμπλέκονται σε μια ασθένεια και των τιμών έκφρασης σε όλους τους ιστούς (Εικόνα 6.14). Επιπλέον, εκτιμάται μια αύξηση 23% στην πιθανότητα ανάπτυξης μιας ασθένειας για κάθε 10 επιπλέον αλληλεπιδράσεις (PPIs) για ένα δεδομένο ζεύγος γονιδίων και μείωση κατά 15% για κάθε 10 επιπλέον εξερχόμενες και εισερχόμενες ακμές των ζευγών γονιδίων. Τα φάρμακα που στοχεύουν στοιχεία του δικτύου των GPCR δεν βρέθηκε να μεταβάλλουν τις πιθανότητες εμφάνισης μιας ασθένειας. Αξιοσημείωτο είναι ότι το μοντέλο μας έδειξε μία πιθανότητα ανάπτυξης μιας ασθένειας μεγαλύτερη από 0.8 σε 34 αλληλεπιδράσεις μεταξύ των ζευγών γονίδιων που δεν έχουν βρεθεί να συσχετίζονται με ασθένειες μέχρι στιγμής. Η G-πρωτεΐνη GNAQ συμμετέχει σε 11 από τις περιπτώσεις αυτές, γεγονός που υποδηλώνει ότι ο ρόλος αυτής της πρωτεΐνης σε ασθένειες θα πρέπει να διερευνηθεί περαιτέρω. Τα δεδομένα αυτά, είναι ιδιαίτερα σημαντικά, καθώς έρχονται να επιβεβαιώσουν παλιότερες μελέτες οι οποίες έδειξαν ότι γονίδια με μεγαλύτερη έκφραση (υπό κανονικές συνθήκες), είναι αυτά τα οποία είναι και πιο πιθανό να εμφανίσουν πολυμορφισμούς που θα προκαλούν ασθένεια (Gorlov et al., 2009), ενώ επίσης, πρωτεΐνες με περισσότερες αλληλεπιδράσεις είναι πιθανό να εμπλέκονται σε ασθένεια σε μεγαλύτερο βαθμό από το αναμενόμενο (Jeong et al., 2001).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114



Εικόνα 6.14. Προσαρμοσμένη πρόβλεψη της πιθανότητας εμφάνισης της ασθένειας μαζί με τα 95% διαστήματα εμπιστοσύνης ανά ιστό για τους τρεις τύπους των ζευγών γονιδίων και του γινομένου της τιμής έκφρασης των αλληλεπιδρώντων γονιδίων (λογαριθμική κλίμακα) που προέρχεται από το μοντέλο λογιστικής παλινδρόμησης.

6.4 Συμπεράσματα

Αυτή είναι η πρώτη μελέτη του δικτύου σηματοδότησης των ανθρώπινων GPCR. Αρχικά δημιουργήθηκε, ένα ενιαίο δίκτυο που περιλαμβάνει τις λειτουργικές σχέσεις μεταξύ των συνιστωσών του συστήματος σηματοδότησης των GPCR. Η ανάλυση του συνολικού δικτύου δεν επιτρέπει την ανίχνευση συγκεκριμένων λειτουργιών που συμβαίνουν σε συγκεκριμένους ιστούς και για το λόγο αυτό χρησιμοποιήθηκαν πληροφορίες σχετικά με το προφίλ έκφρασης ειδικής ανά ιστό. Έτσι δημιουργήθηκαν 11 επιπλέον δίκτυα ειδικά για κάθε ιστό. Το συνολικό δίκτυο και τα 11 ανά ιστό δίκτυα έχουν τις ίδιες τοπολογικές ιδιότητες κάτι το οποίο οδηγεί στο συμπέρασμα ότι ο τρόπος μετάδοσης των εξωκυτταρικών σημάτων μέσω του προσδέτη <GPCR <G-πρωτεΐνη<τελεστή είναι ίδιος. Τα δίκτυα αυτά είναι άνευ κλίμακας, γεγονός που υποδηλώνει ότι κυριαρχούν λίγοι κόμβοι με μεγάλο αριθμό αλληλεπιδράσεων. Δίκτυα άνευ κλίμακας συναντώνται συχνά στα βιολογικά συστήματα όπως

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

είναι τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων (protein-protein interactions networks) (Nacher et al., 2009), τα μεταβολικά και τα βιοχημικά δίκτυα (Jeong et al., 2000).

Οι γενετικές ασθένειες εμφανίζουν παθολογικές εκδηλώσεις που συχνά περιορίζονται σε συγκεκριμένους ιστούς και, ως εκ τούτου, τα γονίδια τα οποία είναι υπεύθυνα για αυτές τις ασθένειες μπορεί να έχουν ποικίλες επιπτώσεις στους διαφορετικούς ιστούς (Guan et al., 2012; Lage et al., 2007). Σε αυτήν τη μελέτη, η έκφραση του γονιδίου σε ιστούς που επηρεάζονται από μια ασθένεια ήταν υψηλότερη σε σύγκριση με τα γονίδια σε ιστούς οι οποίοι δε σχετίζονται με κάποια ασθένεια. Αυτό επιβεβαιώνει προηγούμενες μελέτες που έδειξαν ότι τα γονίδια με τα υψηλότερα επίπεδα έκφρασης κάτω από κανονικές συνθήκες είναι αυτά τα οποία είναι πιο πιθανό να έχουν πολυμορφισμούς που προδιαθέτουν την εμφάνιση της ασθένειας (Gorlov et al., 2009) και ότι οι πρωτεΐνες με περισσότερους αλληλεπιδράσεις είναι πιο πιθανό να ενοχοποιηθούν για την εμπλοκή τους σε ασθένειες (Jeong et al., 2001). Για την παροχή περαιτέρω υποστήριξης στην προαναφερθείσα υπόθεση, η G-πρωτεΐνη GNAI2 η οποία συσχετίζεται με διάφορες ασθένειες ήταν μεταξύ εκείνων με τον μεγαλύτερο αριθμό των αλληλεπιδράσεων.

Θεωρείται ότι τα περισσότερα γονίδια ασκούν τις λειτουργίες τους με τη συμμετοχή τους στο ίδιο σύμπλοκο (Lage et al., 2007). Ως εκ τούτου, η παρουσία των ζευγών των γονιδίων που εκφράζονται στον ίδιο ιστό υποδεικνύει ότι αυτά τα δύο γονίδια εμπλέκονται και στην ίδια κυτταρική διαδικασία. Προκειμένου να διερευνηθεί αυτή η υπόθεση, μελετήθηκαν τα προφίλ έκφρασης των ζευγών γονιδίων και παρατηρήθηκαν ενιαία μοτίβα έκφρασης των ζευγών γονιδίων σε όλους τους ιστούς. Για παράδειγμα, η έκφραση του ζεύγους G-πρωτεΐνη-τελεστής βρέθηκε να είναι υψηλότερη σε σύγκριση με το ζεύγος GPCR-G-πρωτεΐνη σε όλους τους ιστούς το οποίο εμφανίζει συνέπεια σε σχέση με τη μετάδοση του σήματος κατά μήκος του μονοπατιού σηματοδότησης των GPCR.

Επιπλέον, αποδείχθηκε ότι 916 ζεύγη γονιδίων συσχετίζονται με μεγάλη ποικιλία ασθενειών ή κατηγορίες ασθενειών τα οποία αντιπροσωπεύουν το 55% του συνόλου των αλληλεπιδράσεων γονιδίου-γονιδίου. Η έκφραση των γονιδίων που συσχετίζονται με ασθένειες φαίνεται να είναι σταθερά υψηλότερη σε σύγκριση με εκείνη των αντίστοιχων ζευγών γονίδιων που δεν εμπλέκονται σε ασθένεια τα οποία μπορούν να αποτελέσουν ισχυρούς προγνωστικούς παράγοντες για τον κίνδυνο ανάπτυξης μιας ασθένειας. Ωστόσο, αποδείχθηκε οτι τα ζεύγη γονίδιων που εκφράζονται στον υποθάλαμο δε μπορούν να αποτελέσουν προγνωστικούς παράγοντες για τον κίνδυνο εμφάνισης μιας ασθένειας. Αυτό αποδεικνύεται και βιολογικά γιατί ο υποθάλαμος εκκρίνει διαφορετικές ορμόνες που διεγείρουν ή αναστέλλουν την παραγωγή άλλων ορμονών μέσα στους ιστούς σε όλο το σώμα (Melmed et al., 2011).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Στο σύνολό τους, τα παραπάνω ευρήματα υπογραμμίζουν τη σχέση μεταξύ γονιδίων που συσχετίζονται με μια ασθένεια, συν-εκφρασμένων γονιδίων και γονιδίων ειδικών ανά ιστό. Επιπλέον, το μοντέλο παλινδρόμησης που δημιουργήθηκε σε αυτή τη μελέτη μπορεί να χρησιμοποιηθεί για να προβλέψει την πιθανότητα ανάπτυξης μιας ασθένειας με ένα γρήγορο και αξιόπιστο τρόπο. Ως εκ τούτου, η απόκτηση λειτουργικών πληροφοριών για συγκεκριμένους ιστούς που εμφανίζουν συγκεκριμένες (ειδικές ανά ιστό/tissue specific) ασθένειες είναι πολύ σημαντική στον προσδιορισμό των βιολογικών δεικτών για τη διάγνωση, την πρόγνωση και την παρακολούθηση της ασθένειας. Τα ειδικά ανά ιστό δίκτυα αποκάλυψαν μοριακά μονοπάτια τα οποία συμμετέχουν στην ίδια ασθένεια ή κατηγορία ασθένειας κάτι το οποίο δε μπορεί να ανιχνευθεί στο συνολικό δίκτυο. Επίσης, βρέθηκαν λίγα μονοπάτια με εμπλεκόμενες ασθένειες σε συγκεκριμένους ιστούς. Αυτό πιθανώς να οφείλεται στο γεγονός ότι τα μονοπάτια μεταγωγής σήματος διασχίζουν πολλαπλούς ιστούς (Guan et al., 2012). Ένας περιορισμός αυτής της μελέτης ήταν η χαμηλή διαθεσιμότητα των δεδομένων γονιδιακής έκφρασης ειδικής ανά ιστό, καθώς είναι δυνατό μέσα σε κάθε ιστό να υπάρχουν διαφοροποιήσεις στην έκφραση σε κύτταρα διαφόρων τύπων. Επιπλέον, τα γονίδια που κωδικοποιούν τα υπό μελέτη μόρια αποκτήθηκαν από τη βάση δεδομένων Human-gpDB που περιέχει αλληλουχίες πεπτιδίων και πρωτεϊνών. Συνεπώς, υπάρχουν και άλλοι τύποι μορίων (όπως νευροδιαβιβαστές, ιόντα, φωτόνια, κ.λπ.) για τα οποία οι πληροφορίες δεν περιλήφθηκαν σε αυτήν την μελέτη. Παρ' όλα αυτά, σε αυτή τη μελέτη πραγματοποιήθηκε μια πρώτη προσέγγιση βασισμένη στη δημιουργία ενός δικτύου με σκοπό τη διερεύνηση των σχέσεων και των αλληλεπιδράσεων μεταξύ των συνιστωσών του δικτύου σηματοδότησης των GPCR, τόσο σε φυσιολογικές όσο και σε παθολογικές καταστάσεις.

Κλείνοντας πρέπει να αναφερθεί οτι η παραπάνω μελέτη έχει υποβληθεί σε διεθνές επιστημονικό περιοδικό με τίτλο: "***The human GPCR signal transduction network.”*** (Kontou et al., 2016e).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Κεφάλαιο 7

Συζήτηση και Συμπεράσματα

Η παρούσα διδακτορική διατριβή ασχολήθηκε με ένα εύρος θεμάτων και ερευνητικών ερωτημάτων της βιοπληροφορικής, τα οποία είχαν κοινό παρονομαστή τη μεθοδολογία υπολογιστικής ανάλυσης βιολογικών δεδομένων και την προσπάθεια εντοπισμού των παραγόντων που εμπλέκονται στην αιτιολογία πολυπαραγοντικών ασθενειών. Ειδικότερα, χρησιμοποιήθηκαν δεδομένα από γονιδιακούς πολυμορφισμούς (SNPs) και η συσχέτισή τους με ασθένειες, δεδομένα γονιδιακής έκφρασης και ο τρόπος που αυτά επηρεάζουν τις ασθένειες, και τέλος, πραγματοποιήθηκε ανάλυση πολύπλοκων βιολογικών δικτύων που αφορούν τη μοριακή βάση ασθενειών, ενσωματώνοντας δεδομένα από πολλαπλές πηγές.

Στο Κεφάλαιο 2 της διατριβής χρησμοποιήθηκαν μεθοδολογίες μέτα-ανάλυσης γενετικών δεδομένων για τη διερεύνηση της συσχέτισης γονιδιακών πολυμορφισμών με ασθένειες. Όπως αναφέρθηκε ήδη, η τεχνολογική εξέλιξη οδήγησε στη ραγδαία παραγωγή γενετικών δεδομένων. Για πολλές γενετικές συσχετίσεις υπάρχει μεγάλος αριθμός μελετών οι οποίες όμως σε πολλές περιπτώσεις αποκλίνουν και έτσι δεν μπορεί να εξαχθεί ένα τελικό συμπέρασμα αναφορικά με τη συσχέτιση του πολυμορφισμού. Για το λόγο αυτό έχει επικρατήσει η μεθοδολογία της μετα-ανάλυσης η οποία συνδυάζει τα αποτελέσματα των επιμέρους μελετών που ερευνούν την ίδια συσχέτιση προκειμένου να εξαχθεί ένα τελικό αποτέλεσμα. Στην παρούσα διατριβή πραγματοποιήθηκαν τρεις μετα-αναλύσεις γενετικών δεδομένων. Η πρώτη μετα-ανάλυση είχε ως στόχο τη διερεύνηση των πολυμορφισμών του γονιδίου CD24 με τον κίνδυνο εμφάνισης της σκλήρυνσης κατά πλάκας. Πραγματοποιήθηκε μετα-ανάλυση σε τέσσερις πολυμορφισμούς του γονιδίου και για δυο από αυτούς βρέθηκε στατιστικά σημαντική συσχέτιση με την ασθένεια. Στη συνέχεια, πραγματοποιήθηκε μετα-ανάλυση των πολυμορφισμών του γονιδίου FZD3 με τον κίνδυνο εμφάνισης της σχιζοφρένειας. Συνολικά διερευνήθηκαν έξι πολυμορφισμοί. Για τους τέσσερις εκ των οποίων δεν βρέθηκε κάποια γενετική συσχέτιση με την ασθένεια ενώ για έναν πολυμορφισμό βρέθηκε ότι έχει προστατευτικό ρόλο για την εμφάνιση της νόσου. Τέλος, πραγματοποιήθηκε μια τρίτη μετα-ανάλυση με στόχο τη διερεύνηση του ρόλου των γενετικών πολυμορφισμών του συστήματος ρενίνης-αγγειοτενσίνης σε νεφρικές ασθένειες. Εξετάστηκαν δυο πολυμορφισμοί των γονιδίων AGTR1 και AGTR2 σε τέσσερις διαφορετικές κατηγορίες νεφρικών ασθενειών (Χρόνια Νεφρική Νόσος, Νεφρική Νόσος Τελικού Σταδίου,

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Νεφροπάθεια IgA (IgAN) και Κυστεοουρητηρική Παλινδρόμηση) και δεν βρέθηκε κάποια στατιστικά σημαντική συσχέτιση των πολυμορφισμών με τις προαναφερθείσες ασθένειες.

Στο Κεφάλαιο 3 και 4 της διατριβής χρησιμοποιήθηκαν δεδομένα γονιδιακής έκφρασης για τη διευρένηση του τρόπο που αυτά εμπλέκονται σε ασθένειες. Αρχικά, πραγματοποιήθηκε μια εκτενής ανασκόπηση των στατιστικών μεθόδων που έχουν προταθεί για τον εντοπισμό διαφορικώς εκφρασμένων γονιδίων από μικροσυστοιχίες DNA και στη συνέχεια υλοποιήθηκαν στο στατιστικό πακέτο STATA δυο στατιστικές μέθοδοι ανάλυσης δεδομένων μικροσυστοιχιών βασισμένες στον έλεγχο t-test. Η υλοποίηση των μεθόδων είναι διαθέσιμη στο ευρύ κοινό στην ιστοσελίδα: http://www.compgen.org/tools/microarrays#TOC-Methods-of-analysis. Στο Κεφάλαιο 4 πραγματοποιήθηκε ανασκόπηση σχετικά με τη μεθοδολογία που ακολουθείται για τη μετα-ανάλυση δεδομένων γονιδιακής έκφρασης και υλοποιήθηκαν δυο μέθοδοι μετα-ανάλυσης βασισμένες στον έλεγχο t-test, ενώ προτάθηκε και μια καινούργια παραλλαγή της μεθόδου η οποία εμφανίζει μια σειρά από πλεονεκτήματα. Η υλοποίηση των μεθόδων είναι διαθέσιμη στο ευρύ κοινό στην ιστοσελίδα: http://www.compgen.org/tools/microarrays#TOC-Methods-of-meta-analysis. Επίσης προτάθηκαν και άλλοι πιθανοί τρόποι αντιμετώπισης παρόμοιων θεμάτων, οι οποίοι θα μπορούσαν να διερευνηθούν στο μέλλον. Τέλος, οι παραπάνω μέθοδοι εφαρμόστηκαν προκειμένου να βρεθούν γονίδια που εμπλέκονται στο έμφραγμα του μυοκαρδίου. Πραγματοποιήθηκε μετα-ανάλυση σε 31180 γονίδια από τέσσερις μελέτες μικροσυστοιχιών DNA προκειμένου να βρεθούν αυτά που υπερ- ή υπο-εκφράζονται στο έμφραγμα του μυοκαρδίου. Εντοπίστηκαν συνολικά 626 γονίδια τα οποία εμφανίζουν στατιστικώς σημαντικές διαφορές στην έκφραση, ενώ λίγα από αυτά θα είχαν εντοπιστεί χωρίς τη μεθοδολογία της μέτα-ανάλυσης. Για τα 88 από αυτά βρέθηκε ότι εμπλέκονται σε ένα εκτεταμένο δίκτυο πρωτεϊνικών αλληλεπιδράσεων, στο οποίο μια μεγάλη ομάδα απαρτίζεται από κυτοκίνες και υποδοχείς αυτών, μια άλλη ομάδα από γονίδια που σχετίζονται με τη μεταφορά RNA, ενώ γονίδια όπως της SERPIN βρέθηκαν να έχουν κεντρικό ρόλο. Αξίζει τέλος να σημειωθεί, ότι από τα 626 γονίδια, μόνο 7 βρέθηκαν από τη βιβλιογραφία να εμπλέκονται και σε γενετική προδιάθεση για την ασθένεια.

Στο Κεφάλαιο 5, κατασκευάστηκε ένα ενοποιημένο μοντέλο που χρησιμοποιεί δεδομένα από διαφορετικές πηγές προκειμένου να εξαχθεί ένα συμπέρασμα σχετικά με την εμπλοκή των γονιδίων σε ασθένειες. Χρησιμοποιώντας αξιόπιστα δεδομένα γενετικής συσχέτισης από διαφορετικές πηγές, δημιουργήθηκε ένα δίκτυο αλληλεπιδράσεων ανθρώπινων γονιδίων και ασθενειών, προκειμένου να διερευνηθούν οι συσχετίσεις μεταξύ των γενετικών ασθενειών του ανθρώπου και των γονιδίων που σχετίζονται με τις ασθένειες αυτές. Η ανάλυση αυτή έδειξε την ανάγκη ύπαρξης αξιόπιστων και

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

συνεχώς ανανεούμενων βάσεων δεδομένων, κυρίως για τις περιπτώσεις χαμηλής διεισδυτικότητας, ανάγκη που σήμερα είναι περισσότερο επιτακτική λόγω της διακοπής λειτουργίας της βάσης δεδομένων GAD. Στο ίδιο πλαίσιο, έγινε αντιληπτή και η ανάγκη τυποποίησης των ονομάτων των γονιδίων αλλά και αυτών των ασθενειών, καθώς διαφορετικές βάσεις δεδομένων χρησιμοποιούν διαφορετικές συμβάσεις. Στη συνέχεια με βάση αυτό το δίκτυο δημιουργήθηκε ένα δίκτυο γονιδίων και ένα δίκτυο ασθενειών με στόχο την εύρεση ασθενειών που προκαλούνται από τα ίδια γονίδια, αλλά και γονιδίων που εμπλέκονται στην ίδια ασθένεια. Η ανάλυση των δικτύων αυτών και η εφαρμογή της θεωρίας των γράφων και των πολύπλοκων συστημάτων, έδωσε αρκετά σημαντικά συμπεράσματα σχετικά με τη φύση των πολυπαραγοντικών ασθενειών και την εμπλοκή των γονιδίων σε αυτές. Το δίκτυο γονιδίων-γονιδίων αποτελεί μια απεικόνιση της αρχιτεκτονικής των ασθενειών που εμφανίζονται στον άνθρωπο (human diseasome) και θα μπορούσε να είναι ιδιαίτερα χρήσιμο για τον εντοπισμό των γενετικών παραγόντων που συμβάλλουν στην αιτιολογία, στην προδιάθεση (susceptibility) και στη συννοσηρότητα (comorbitity) των ασθενειών. Τα αποτελέσματα αυτά μπορεί να έχουν σημαντική εφαρμογή στη διάγνωση, την πρόληψη και τη θεραπεία ασθενειών, καθώς και στο σχεδιασμό φαρμάκων. Τα φαρμακολογικά δίκτυα αποτελούν ένα απαραίτητο εργαλείο για την ανάπτυξη νέων θεραπευτικών στρατηγικών (Berger and Iyengar, 2009; Hopkins, 2009; Wang et al., 2012b). Μια λεπτομερής κατανόηση του πώς συνδέονται φαινομενικά διαφορετικές ασθένειες είναι απαραίτητη για την ανάπτυξη αποτελεσματικών θεραπειών. Μέχρι στιγμής, οι προσπάθειες έχουν στραφεί κυρίως προς την αναζήτηση θεραπευτικών ουσιών που στοχεύουν τα γονίδια που συμμετέχουν σε συγκεκριμένες ασθένειες. Με τον τρόπο αυτό, οι ασθένειες που συνδέονται με γονίδια με γνωστή συσχέτιση αντιμετωπίζονται, ενώ άλλες ασθένειες με παρόμοιες κλινικές εκδηλώσεις δεν αντιμετωπίζονται (Kuhn et al., 2010; Nolan, 2007). Ως εκ τούτου, δύο ή περισσότερα φαινοτυπικά ανεξάρτητες ασθένειες με την ίδια γενετική προέλευση θα μπορούσαν να θεραπευθούν μέσω της δράσης ενός μόνο φαρμάκου. Οι σημαντικές συσχετίσεις ασθενειών–ασθενειών που ανιχνεύθηκαν με κοινή γενετική βάση μπορούν να αποτελέσουν το θεμέλιο λίθο, στον οποίο θα στηριχθούν οι μελλοντικές προσπάθειες. Τα δεδομένα που προέκυψαν από την παραπάνω ανάλυση είναι δημόσια διαθέσιμα στην αντίστοιχη δημοσίευση στο περιοδικό *Data in Brief*.

Τέλος, στο Κεφάλαιο 6, κατασκευάστηκε και μελετήθηκε για πρώτη φορά στη βιβλιογραφία το δίκτυο μεταγωγής σήματος των ανθρώπινων υποδοχέων συζευγμένων με G-πρωτεΐνες (GPCRs) ώστε να διερευνηθεί η πιθανή συσχέτιση της δράσης των υποδοχέων αυτών με τον κίνδυνο εμφάνισης ασθενειών. Τα γονίδια που κωδικοποιούν τους GPCRs έχουν τεράστια φαρμακολογική σημασία, με

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

περίπου το 40% όλων των φαρμάκων που κυκλοφορούν στην αγορά να στοχεύουν αυτές τις πρωτεΐνες (Hopkins and Groom, 2002; Overington et al., 2006). Τα μονοπάτια μεταγωγής σήματος των GPCR εμπλέκονται σε μια πληθώρα ασθενειών (Dorsam and Gutkind, 2007; Thompson et al., 2014; Vischer et al., 2014). Ως εκ τούτου, η αποσαφήνιση των μοριακών μηχανισμών του μονοπατιού της μεταγωγής σήματος των GPCR καθώς και ο τρόπος με τον οποίο το μονοπάτι αυτό συνδέεται με ασθένειες είναι υψίστης βιολογικής και φαρμακευτικής σημασίας. Η ανάλυση αυτού του δικτύου αλλά και η ενοποίηση διαφορετικών τύπων δεδομένων στην ανάλυση, έδωσε σημαντικές πληροφορίες για τον τρόπο δράσης των υποδοχέων αυτών αλλά και για τον τρόπο που αυτοί επηρεάζουν την εμφάνιση γνωστών ασθενειών. Για παράδειγμα. το δίκτυο βρέθηκε ότι ανήκει στην κατηγορία των δικτύων άνευ κλίμακας, και μάλιστα έχει τα ίδια γενικά χαρακτηριστικά ανεξαρτήτως του ιστού. Με άλλα λόγια, οι γενικές ιδιότητες του συστήματος μεταγωγής σήματος είναι ίδιες σε κάθε ιστό, παρόλο που συμμετέχουν κάθε φορά διαφορετικά γονίδια. Βρέθηκε επίσης ότι η γονιδιακή έκφραση σε ιστούς που επηρεάζονται από μια ασθένεια ήταν υψηλότερη σε σύγκριση με τα γονίδια σε ιστούς οι οποίοι δε σχετίζονται με κάποια ασθένεια. Αυτό επιβεβαιώνει προηγούμενες μελέτες που έδειξαν ότι τα γονίδια με τα υψηλότερα επίπεδα έκφρασης κάτω από κανονικές συνθήκες είναι αυτά τα οποία είναι πιο πιθανό να έχουν πολυμορφισμούς που προδιαθέτουν την εμφάνιση της ασθένειας (Gorlov et al., 2009) και ότι οι πρωτεΐνες με περισσότερες αλληλεπιδράσεις είναι πιο πιθανό να ενοχοποιηθούν για την εμπλοκή τους σε ασθένειες (Jeong et al., 2001), ενώ τα ειδικά ανά ιστό δίκτυα αποκάλυψαν μοριακά μονοπάτια τα οποία συμμετέχουν στην ίδια ασθένεια ή κατηγορία ασθένειας κάτι το οποίο δε μπορεί να ανιχνευθεί στο συνολικό δίκτυο.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Βιβλιογραφία

Alberti, K.G., Zimmet, P., Shaw, J., and Group, I.D.F.E.T.F.C. (2005). The metabolic syndrome--a new worldwide definition. Lancet *366*, 1059-1062.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (1994). Molecular Biology of the Cell, 3rd edn (Garland Publishing, Inc).

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic acids research *43*, D789-798.

Andrews, D.W., and Buchinsky, M. (2000). A Three-step Method for Choosing the Number of Bootstrap Repetitions. Econometrica *68*, 23-51.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-230.

Attard-Montalto, S.P., Saha, V., Ng, Y.Y., Kingston, J.E., and Eden, O.B. (1994). High incidence of hypertension in children presenting with acute lymphoblastic leukemia. Pediatric hematology and oncology *11*, 519-525.

Attwood, T.K., and Findlay, J.B. (1994). Fingerprinting G-protein-coupled receptors. Protein engineering *7*, 195-203.

Bagos, P.G. (2008). A unification of multivariate methods for meta-analysis of genetic association studies. Statistical applications in genetics and molecular biology *7*, Article31.

Bagos, P.G., Dimou, N.L., Liakopoulos, T.D., and Nikolopoulos, G.K. (2011). Meta-analysis of family-based and case-control genetic association studies that use the same cases. Statistical Applications in Genetics and Molecular Biology *10*.

Bagos, P.G., Dimou, Niki L, Liakopoulos, Theodore D, & Nikolopoulos, Georgios K (2011). Meta-analysis of family-based and case-control genetic association studies that use the same cases. Statistical applications in genetics and molecular biology *10*, 1-41.

Bagos, P.G., and Nikolopoulos, G.K. (2009). Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. Journal of clinical epidemiology *62*, 1037-1044.

Bai, X.F., Li, O., Zhou, Q., Zhang, H., Joshi, P.S., Zheng, X., Liu, Y., Wang, Y., Zheng, P., and Liu, Y. (2004). CD24 controls expansion and persistence of autoreactive T cells in the central nervous system during experimental autoimmune encephalomyelitis. The Journal of experimental medicine *200*, 447-458.

Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48-54.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2005). The Universal Protein Resource (UniProt). Nucleic acids research *33 Database Issue*, D154-159.

Baldi, P., and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics *17*, 509-519.

Barabasi, A.L. (2009). Scale-free networks: a decade and beyond. Science *325*, 412-413.

Barabasi, A.L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509-512.

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews *12*, 56-68.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviews Genetics *5*, 101-113.

Barrenas, F., Chavali, S., Holme, P., Mobini, R., and Benson, M. (2009). Network properties of complex human disease genes identified through genome-wide association studies. PloS one *4*, e8090.

Barrett, T., and Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. Methods in molecular biology (Clifton, NJ *338*, 175-190.

Basset el, E.A., Berthoux, P., Cecillon, S., Deprle, C., Thibaudin, D., De Filippis, J.P., Alamartin, E., and Berthou, F. (2002). Hypertension after renal transplantation and polymorphism of genes involved in essential hypertension: ACE, AGT, AT1 R and ecNOS. Clinical nephrology *57*, 192-200.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M.A., Sanz, F., and Furlong, L.I. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PloS one *6*, e20284.

Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The genetic association database. Nat Genet *36*, 431-432.

Begg, C.B., and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics *50*, 1088-1101.

Benjafield, A.V., Katyk, K., and Morris, B.J. (2003). Association of EDNRA, but not WNK4 or FKBP1B, polymorphisms with essential hypertension. Clinical genetics *64*, 433-438.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Benjamini, Y., Krieger, A.M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika *93*, 491-507.

Benjamini, Y., and Liu, W. (1999a). A distribution-free multiple test procedure that controls the false discovery rate. Tel Aviv, Tel Aviv University.

Benjamini, Y., and Liu, W. (1999b). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference *82*, 163-170.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.

Bentall, R.P. (2013). Reconstructing schizophrenia (Routledge).

Berger, S.I., and Iyengar, R. (2009). Network analyses in systems pharmacology. Bioinformatics *25*, 2466-2472.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., *et al.* (2002). The Protein Data Bank. Acta Crystallogr D Biol Crystallogr *58*, 899-907.

Bonaventura, A., Montecucco, F., and Dallegri, F. (2016). Cellular recruitment in myocardial ischaemia/reperfusion injury. European journal of clinical investigation *46*, 590-601.

Braliou, G.G., Grigoriadou, A.M., Kontou, P.I., and Bagos, P.G. (2014). The role of genetic polymorphisms of the Renin-Angiotensin System in renal diseases: A meta-analysis. Computational and structural biotechnology journal *10*, 1-7.

Braliou, G.G., Pantavou, K.G., Kontou, P.I., and Bagos, P.G. (2015). Polymorphisms of the CD24 Gene Are Associated with Risk of Multiple Sclerosis: A Meta-Analysis. International journal of molecular sciences *16*, 12368-12381.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., *et al.* (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic acids research *31*, 68-71.

Brazma, A., and Vilo, J. (2000). Gene expression data analysis. FEBS letters *480*, 17-24.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters *573*, 83-92.

Breitling, R., and Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. Journal of Bioinformatics and Computational Biology *3*, 1171-1189.

Cabrera-Vera, T.M., Vanhauwe, J., Thomas, T.O., Medkova, M., Preininger, A., Mazzoni, M.R., and Hamm, H.E. (2003). Insights into G protein structure, function, and regulation. Endocrine reviews *24*, 765-781.

Campain, A., and Yang, Y.H. (2010). Comparison study of microarray meta-analysis methods. BMC bioinformatics *11*, 408.

Chang, L.C., Lin, H.M., Sibille, E., and Tseng, G.C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics *14*, 368.

Choi, J.K., Yu, U., Kim, S., and Yoo, O.J. (2003). Combining multiple microarray studies and modeling interstudy variation. Bioinformatics (Oxford, England) *19 Suppl 1*, i84-90.

Chou, P.Y., and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol *47*, 45-148.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

ChuanYuan, K., Li, Z., Hua, L., and JianZhong, Y. (2011). Association study of the frizzled 3 gene with Chinese Va schizophrenia. Neuroscience letters *505*, 196-199.
Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. Nature protocols *2*, 2366-2382.
Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L (Erlbaum).
Conklin, B.R., and Bourne, H.R. (1993). Structural elements of G alpha subunits that interact with G beta gamma, receptors, and effectors. Cell *73*, 631-641.
Conlon, E.M., Song, J.J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. BMC bioinformatics *8*, 80.
Conlon, E.M., Song, J.J., and Liu, J.S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. BMC bioinformatics *7*, 247.
Consortium, M., Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., *et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature biotechnology *24*, 1151-1161.
Cordell, H.J., and Clayton, D.G. (2005). Genetic association studies. Lancet *366*, 1121-1131.
Cousins, R.D. (2007). Annotated bibliography of some papers on combining significances or p-values. arXiv preprint arXiv:07052209.
Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138-163.
Crick, F.H., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. Nature *192*, 1227-1232.
Davidson, R., and MacKinnon, J.G. (2000). Bootstrap tests: How many bootstraps? Econometric Reviews *19*, 55-68.
Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J.C., Nitzberg, M., Wymore, F., Zachariah, Z.K., *et al.* (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic acids research *35*, D766-770.
DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials *7*, 177-188.
Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., and Chinnaiyan, A.M. (2001). Delineation of prognostic biomarkers in prostate cancer. Nature *412*, 822-826.
Dorsam, R.T., and Gutkind, J.S. (2007). G-protein-coupled receptors and cancer. Nature reviews Cancer *7*, 79-94.
Dudoit, S.Y.H.Y., Matthew J. Callow, and Terence P. Speed (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report # 578.
Ebers, G.C. (2008). Environmental factors and multiple sclerosis. The Lancet Neurology *7*, 268-277.
Ebers, G.C., Sadovnick, A.D., and Risch, N.J. (1995). A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. Nature *377*, 150-151.
Edgington, E.S. (1972). An additive method for combining probability values from independent experiments. The Journal of Psychology *80*, 351-363.
Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, Vol 38 (SIAM).
Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association *82*, 171-185.
Efron, B., and Tibshirani, R. (1993). An Introduction to the Bootstrap (Boca Raton, FL: Chapman & Hall/CRC).
Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association *96*, 1151-1160.
Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ (Clinical research ed *315*, 629-634.
Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research *30*, 1575-1584.
Fisher, R.A. (1946). Statistical methods for research workers. Statistical methods for research workers.
Fox, R.J., and Dimmic, M.W. (2006). A two-sample Bayesian t-test for microarray data. BMC bioinformatics *7*, 126.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Freeman, W.M., Robertson, D.J., and Vrana, K.E. (2000). Fundamentals of DNA hybridization arrays for gene expression analysis. BioTechniques *29*, 1042-1046, 1048-1055.
Friedrich, J.O., Adhikari, N.K., and Beyene, J. (2008). The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Medical Research Methodology *8*, 1.
Gether, U. (2000). Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocrine reviews *21*, 90-113.
Gobbi, A., Iorio, F., Dawson, K.J., Wedge, D.C., Tamborero, D., Alexandrov, L.B., Lopez-Bigas, N., Garnett, M.J., Jurman, G., and Saez-Rodriguez, J. (2014). Fast randomization of large genomic datasets while preserving alteration counts. Bioinformatics *30*, i617-623.
Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proceedings of the National Academy of Sciences of the United States of America *104*, 8685-8690.
Goldstein, D.B. (2009). Common genetic variation and human traits. The New England journal of medicine *360*, 1696-1698.
Gonen, M. (2010). The Bayesian t-test and beyond. Methods Mol Biol *620*, 179-199.
Gönen, M., Johnson, W.O., Lu, Y., and Westfall, P.H. (2005). The Bayesian two-sample t test. The American Statistician *59*, 252-257.
Gorlov, I.P., Gallick, G.E., Gorlova, O.Y., Amos, C., and Logothetis, C.J. (2009). GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. PloS one *4*, e6511.
Gottardo, R., Pannucci, J.A., Kuske, C.R., and Brettin, T. (2003). Statistical analysis of microarray data: a Bayesian approach. Biostatistics *4*, 597-620.
Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org: the HGNC resources in 2015. Nucleic acids research *43*, D1079-1085.
Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A.K., Schimenti, J.C., Handel, M.A., Bult, C.J., Hibbs, M.A., and Troyanskaya, O.G. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS computational biology *8*, e1002694.
Gumprecht, J., Zychma, M.J., Grzeszczak, W., and Zukowska-Szczechowska, E. (2000). Angiotensin I-converting enzyme gene insertion/deletion and angiotensinogen M235T polymorphisms: risk of chronic renal failure. End-Stage Renal Disease Study Group. Kidney international *58*, 513-519.
HapMap (2003). The International HapMap Project. Nature *426*, 789-796.
Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47-52.
Hashimoto, R., Suzuki, T., Iwata, N., Yamanouchi, Y., Kitajima, T., Kosuga, A., Tatsumi, M., Ozaki, N., Kamijima, K., and Kunugi, H. (2005). Association study of the frizzled-3 (FZD3) gene with schizophrenia and mood disorders. Journal of neural transmission *112*, 303-307.
Hernandez-Campo, P.M., Almeida, J., Matarraz, S., de Santiago, M., Sanchez, M.L., and Orfao, A. (2007). Quantitative analysis of the expression of glycosylphosphatidylinositol-anchored proteins during the maturation of different hematopoietic cell compartments of normal bone marrow. Cytometry Part B, Clinical cytometry *72*, 34-42.
Heyer, L.J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome research *9*, 1106-1115.
Higgins, J.P., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. Bmj *327*, 557-560.
Hirschhorn, J.N. (2009). Genomewide association studies--illuminating biologic pathways. The New England journal of medicine *360*, 1699-1701.
Holland, B.S., and Copenhaver, M.D. (1987). An Improved Sequentially Rejective Bonferroni Test Procedure. Biometrics *43*, 417-423.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics *6*, 65-70. .

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Hong, F., and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics *24*, 374-382.
Hopkins, A.L. (2009). Drug discovery: Predicting promiscuity. Nature *462*, 167-168.
Hopkins, A.L., and Groom, C.R. (2002). The druggable genome. Nature reviews Drug discovery *1*, 727-730.
Hu, P., Greenwood, C.M., and Beyene, J. (2009). Using the ratio of means as the effect size measure in combining results of microarray experiments. BMC Syst Biol *3*, 106.
Hwang, K.B., Kong, S.W., Greenberg, S.A., and Park, P.J. (2004). Combining gene expression data from different generations of oligonucleotide arrays. BMC Bioinformatics *5*, 159.
Ide, M., Muratake, T., Yamada, K., Iwayama-Shigeno, Y., Iwamoto, K., Takao, H., Toyota, T., Kaneko, N., Minabe, Y., and Nakamura, K. (2004). Genetic and expression analyses of< i> FZD3</i> in schizophrenia. Biological psychiatry *56*, 462-465.
International Multiple Sclerosis Genetics, C., Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., *et al.* (2007). Risk alleles for multiple sclerosis identified by a genomewide study. The New England journal of medicine *357*, 851-862.
Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., and Jurman, G. (2009). Repeatability of published microarray gene expression analyses. Nature genetics *41*, 149-155.
Ioannidis, J.P., Boffetta, P., Little, J., O'Brien, T.R., Uitterlinden, A.G., Vineis, P., Balding, D.J., Chokkalingam, A., Dolan, S.M., Flanders, W.D., *et al.* (2008). Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol *37*, 120-132.
Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nature genetics *29*, 306-309.
Ioannidis, J.P., and Trikalinos, T.A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. Journal of clinical epidemiology *58*, 543-549.
Janssen, R., Bont, L., Siezen, C.L., Hodemaekers, H.M., Ermers, M.J., Doornbos, G., van't Slot, R., Wijmenga, C., Goeman, J.J., and Kimpen, J.L. (2007). Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. Journal of Infectious Diseases *196*, 826-834.
Jarvinen, A.K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.P., and Monni, O. (2004). Are data from different gene expression microarray platforms comparable? Genomics *83*, 1164-1168.
Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42.
Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.
Jeong, S.H., Joo, E.J., Ahn, Y.M., Lee, K.Y., and Kim, Y.S. (2006). Investigation of genetic association between human Frizzled homolog 3 gene (< i> FZD3</i>) and schizophrenia: Results in a Korean population and evidence from meta-analysis. Psychiatry Research *143*, 1-11.
Jiang, W., and Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Stat Med *26*, 5320-5334.
Jison, M.L., Munson, P.J., Barb, J.J., Suffredini, A.F., Talwar, S., Logun, C., Raghavachari, N., Beigel, J.H., Shelhamer, J.H., Danner, R.L., *et al.* (2004). Blood mononuclear cell gene expression profiles characterize the oxidant, hemolytic, and inflammatory stress of sickle cell disease. Blood *104*, 270-280.
Kadota, K., Nakai, Y., and Shimizu, K. (2008). A weighted average difference method for detecting differentially expressed genes from microarray data. Algorithms Mol Biol *3*, 8.
Kaiser, J. (2007). An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples. Stata Journal *7*, 402-412.
Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic acids research *44*, D457-462.
Katsu, T., Ujike, H., Nakano, T., Tanaka, Y., Nomura, A., Nakata, K., Takaki, M., Sakai, A., Uchida, N., and Imamura, T. (2003). The human frizzled-3 (< i> FZD3</i>) gene on chromosome 8p21, a receptor gene for Wnt ligands, is associated with the susceptibility to schizophrenia. Neuroscience letters *353*, 53-56.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Kayala, M.A., and Baldi, P. (2012). Cyber-T web server: differential analysis of high-throughput data. Nucleic Acids Res *40*, W553-559.
Kerr, K.F. (2007). Extended analysis of benchmark datasets for Agilent two-color microarrays. BMC bioinformatics *8*, 371.
Kessler, T., Erdmann, J., and Schunkert, H. (2013). Genetics of coronary artery disease and myocardial infarction--2013. Current cardiology reports *15*, 368.
Khomtchouk, B.B., Van Booven, D.J., and Wahlestedt, C. (2014). HeatmapGenerator: high performance RNAseq and microarray visualization software suite to examine differential gene expression levels using an R and C++ hybrid computational pipeline. Source code for biology and medicine *9*, 30.
Kim, J., Ghasemzadeh, N., Eapen, D.J., Chung, N.C., Storey, J.D., Quyyumi, A.A., and Gibson, G. (2014). Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome medicine *6*, 40.
King, N., Hittinger, C.T., and Carroll, S.B. (2003). Evolution of key cell signaling and adhesion protein families predates animal origins. Science *301*, 361-363.
Kolakowski, L.F., Jr. (1994). GCRDb: a G-protein-coupled receptor database. Receptors & channels *2*, 1-7.
Kontou, P.I., Pavlopoulou, A., and Bagos, P.G. (2016a). Methods of Analysis and Meta-Analysis for Identifying Differentially Expressed Genes. Methods in molecular biology (Clifton, NJ.
Kontou, P.I., Pavlopoulou, A., Braliou, G., Bogiatzi, S., and Bagos, P.G. (2016b). Identification of differentially expressed genes in cardiovascular diseases: a Meta-analysis.
Kontou, P.I., Pavlopoulou, A., Dimou, N.L., Pavlopoulos, G.A., and Bagos, P.G. (2016c). Data and programs in support of network analysis of genes and their association with diseases. Data in Brief *8*, 1036-1039.
Kontou, P.I., Pavlopoulou, A., Dimou, N.L., Pavlopoulos, G.A., and Bagos, P.G. (2016d). Network analysis of genes and their association with diseases. Gene *590*, 68-78.
Kontou, P.I., Pavlopoulou, A., Dimou, N.L., Theodoropoulou, M., Braliou, G., Tsaousis, G., Pavlopoulos, G.A., Hamodrakas, S.I., and Bagos, P.G. (2016e). The human GPCR signal transduction network.
Kooperberg, C., Aragaki, A., Strand, A.D., and Olson, J.M. (2005). Significance testing for small microarray experiments. Statistics in medicine *24*, 2281-2298.
Kristiansen, K. (2004). Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. Pharmacology & therapeutics *103*, 21-80.
Krupp, M., Marquardt, J.U., Sahin, U., Galle, P.R., Castle, J., and Teufel, A. (2012). RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics (Oxford, England) *28*, 1184-1185.
Kruschke, J.K. (2013). Bayesian estimation supersedes the t test. J Exp Psychol Gen *142*, 573-603.
Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. Molecular systems biology *6*, 343.
Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N., *et al.* (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology *25*, 309-316.
Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T.C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. The New England journal of medicine *327*, 248-254.
Lau, J., Schmid, C.H., and Chalmers, T.C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. Journal of clinical epidemiology *48*, 45-57; discussion 59-60.
Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., *et al.* (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research *42*, D1091-1097.
Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome research *14*, 1085-1094.
Lewin, B. (1990). Genes IV (Oxford University Press).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Lichtenstein, P., Yip, B.H., Björk, C., Pawitan, Y., Cannon, T.D., Sullivan, P.F., and Hultman, C.M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet *373*, 234-239.

Lifton, R.P., Gharavi, A.G., and Geller, D.S. (2001). Molecular mechanisms of human hypertension. Cell *104*, 545-556.

Liu, C.C., Tseng, Y.T., Li, W., Wu, C.Y., Mayzus, I., Rzhetsky, A., Sun, F., Waterman, M., Chen, J.J., Chaudhary, P.M., *et al.* (2014). DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic acids research *42*, W137-146.

Liu, H., Bebu, I., and Li, X. (2007a). Microarray probes and probe sets. Frontiers in bioscience (Elite edition) *2*, 325-338.

Liu, J.Q., Carl, J.W., Jr., Joshi, P.S., RayChaudhury, A., Pu, X.A., Shi, F.D., and Bai, X.F. (2007b). CD24 on the resident cells of the central nervous system enhances experimental autoimmune encephalomyelitis. Journal of immunology *178*, 6227-6235.

Liu, Y., and Zheng, P. (2007). CD24: a genetic checkpoint in T cell homeostasis and autoimmune diseases. Trends in immunology *28*, 315-320.

Livak, K.J., Marmaro, J., and Todd, J.A. (1995). Towards fully automated genome-wide polymorphism screening. Nat Genet *9*, 341-342.

Lönnstedt, I., and Speed, T. (2002). Replicated microarray data. Statistica sinica, 31-46.

Loughin, T.M. (2004). A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis *47*, 467-485.

Manning, A.K., LaValley, M., Liu, C.T., Rice, K., An, P., Liu, Y., Miljkovic, I., Rasmussen-Torvik, L., Harris, T.B., Province, M.A., *et al.* (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. Genetic epidemiology *35*, 11-18.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. The New England journal of medicine *363*, 166-176.

Marot, G., Foulley, J.L., Mayer, C.D., and Jaffrezic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics *25*, 2692-2699.

McCarthy, D.J., and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics *25*, 765-771.

McCudden, C.R., Hains, M.D., Kimple, R.J., Siderovski, D.P., and Willard, F.S. (2005). G-protein signaling: back to the future. Cellular and molecular life sciences : CMLS *62*, 551-577.

McEntyre, J., and Lipman, D. (2001). PubMed: bridging the information gap. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne *164*, 1317-1319.

McKnight, A.J., Currie, D., and Maxwell, A.P. (2010). Unravelling the genetic basis of renal diseases; from single gene to multifactorial disorders. The Journal of pathology *220*, 198-216.

McPherson, R., and Tybjaerg-Hansen, A. (2016). Genetics of Coronary Artery Disease. Circulation research *118*, 564-578.

Mehta, P.K., and Griendling, K.K. (2007). Angiotensin II cell signaling: physiological and pathological effects in the cardiovascular system. American journal of physiology Cell physiology *292*, C82-97.

Menzaghi, C., Paroni, G., De Bonis, C., Soccio, T., Marucci, A., Bacci, S., and Trischitta, V. (2006). The -318 C>G single-nucleotide polymorphism in GNAI2 gene promoter region impairs transcriptional activity through specific binding of Sp1 transcription factor and is associated with high blood pressure in Caucasians from Italy. Journal of the American Society of Nephrology : JASN *17*, S115-119.

Metzker, M.L. (2010). Sequencing technologies [mdash] the next generation. Nature reviews *11*, 31-46.

Meuwissen, T.H., and Goddard, M.E. (2004). Bootstrapping of gene-expression data improves and controls the false discovery rate of differentially expressed genes. Genet Sel Evol *36*, 191-205.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet *19*, 570-577.

Mortality, G.B.D., and Causes of Death, C. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet *385*, 117-171.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Nacher, J.C., Hayashida, M., and Akutsu, T. (2009). Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs. Bio Systems *95*, 155-159.
Neuhauser, M., and Jockel, K.H. (2006). A bootstrap test for the analysis of microarray experiments with a very small number of replications. Appl Bioinformatics *5*, 173-179.
Nolan, G.P. (2007). What's wrong with drug screening today. Nature chemical biology *3*, 187-191.
Normand, S.L. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. Statistics in medicine *18*, 321-359.
Novoradovskaya, N., Whitfield, M.L., Basehore, L.S., Novoradovsky, A., Pesich, R., Usary, J., Karaca, M., Wong, W.K., Aprelikova, O., Fero, M., *et al.* (2004). Universal Reference RNA as a standard for microarray experiments. BMC genomics *5*, 20.
Nussbaum, R.L., McInnes, R.R., and Willard, H.F. (2007). Thompson & Thompson Genetics in Medicine (Elsevier Health Sciences).
Oda, E., and Kawai, R. (2011). High-density lipoprotein cholesterol is positively associated with hypertension in apparently healthy Japanese men and women. British journal of biomedical science *68*, 29-33.
Oellrich, A., Sanger Mouse Genetics, P., and Smedley, D. (2014). Linking tissues to phenotypes using gene expression profiles. Database : the journal of biological databases and curation *2014*, bau017.
Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R., and Feng, C. (2008). GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update. Nucleic acids research *36*, D907-912.
Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., *et al.* (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic acids research *42*, D358-363.
Oti, M., and Brunner, H.G. (2007). The modular nature of genetic diseases. Clinical genetics *71*, 1-11.
Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nature reviews Drug discovery *5*, 993-996.
Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews *12*, 87-98.
Pantavou, K.G., Braliou, G.G., Kontou, P.I., Dimou, N.L., and Bagos, P.G. (2016). A meta-analysis of FZD3 gene polymorphisms and their association with schizophrenia. Psychiatric genetics *26*, 272-280.
Pare, G., Serre, D., Brisson, D., Anand, S.S., Montpetit, A., Tremblay, G., Engert, J.C., Hudson, T.J., and Gaudet, D. (2007). Genetic analysis of 103 candidate genes for coronary artery disease and associated phenotypes in a founder population reveals a new association between endothelin-1 and high-density lipoprotein cholesterol. American journal of human genetics *80*, 673-682.
Pasquier, C., Promponas, V.J., and Hamodrakas, S.J. (2001). PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. Proteins *44*, 361-369.
Pawson, T., and Linding, R. (2008). Network medicine. FEBS letters *582*, 1266-1270.
Pedersen, L.R., Frestad, D., Michelsen, M.M., Mygind, N.D., Rasmusen, H., Suhrs, H.E., and Prescott, E. (2016). Risk Factors for Myocardial Infarction in Women and Men: A Review of the Current Literature. Current pharmaceutical design.
Pedersen, M.S., Benros, M.E., Agerbo, E., Borglum, A.D., and Mortensen, P.B. (2012). Schizophrenia in patients with atopic disorders with particular emphasis on asthma: a Danish population-based study. Schizophrenia research *138*, 58-62.
Petiti, D.B. (1994). Meta-analysis Decision Analysis and Cost-Effectiveness Analysis, Vol 24 (Oxford University Press).
Pokorny, J., Luan, N.T., Kondratenko, S.S., and Janicek, G. (1976). Changes of sensory value by interaction of alkanals with amino acids and proteins. Nahrung *20*, 267-272.
Pulver, A.E., Lasseter, V.K., Kasch, L., Wolyniec, P., Nestadt, G., Blouin, J.L., Kimberland, M., Babb, R., Vourlis, S., and Chen, H. (1995). Schizophrenia: a genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. American journal of medical genetics *60*, 252-260.
Quackenbush, J. (2001). Computational analysis of microarray data. Nature reviews *2*, 418-427.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

R Core Team (2016). R: A language and environment for statistical computing. In R Foundation for Statistical Computing (Vienna, Austria).
Ramasamy, A., Mondry, A., Holmes, C.C., and Altman, D.G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS medicine *5*, e184.
Reif, A., Melchers, M., Strobel, A., Jacob, C., Herterich, S., Lesch, K.-P., and Zimmer, M. (2007). FZD3 is not a risk gene for schizophrenia: a case-control study in a Caucasian sample (Springer).
Rens-Domiano, S., and Hamm, H.E. (1995). Structural and functional relationships of heterotrimeric G-proteins. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *9*, 1059-1066.
Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.
Rost, B., and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol *232*, 584-599.
Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review *16*, 225-237.
Satagopam, V.P., Theodoropoulou, M.C., Stampolakis, C.K., Pavlopoulos, G.A., Papandreou, N.C., Bagos, P.G., Schneider, R., and Hamodrakas, S.J. (2010). GPCRs, G-proteins, effectors and their interactions: human-gpDB, a database employing visualization tools and data integration techniques. Database : the journal of biological databases and curation *2010*, baq019.
Savva, J., Alfakih, K., Galloway, S.L., Hall, A.S., West, R.M., Ball, S.G., Balmforth, A.J., and Maqbool, A. (2012). The alpha(2C)-Del322-325 adrenoceptor polymorphism and the occurrence of left ventricular hypertrophy in hypertensives. Blood pressure *21*, 116-121.
Schwartz, P.J. (2011). Season of birth in schizophrenia: a maternal-fetal chronobiological hypothesis. Medical hypotheses *76*, 785-793.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.
Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research *29*, 308-311.
Sidak, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association *62*, 626 -633.
Smedley, D., Oellrich, A., Kohler, S., Ruef, B., Sanger Mouse Genetics, P., Westerfield, M., Robinson, P., Lewis, S., and Mungall, C. (2013). PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database : the journal of biological databases and curation *2013*, bat025.
Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (Springer), pp. 397-420.
Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P., Buneman, O.P., Davenport, A.P., McGrath, J.C., Peters, J.A., *et al.* (2016). The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic acids research *44*, D1054-1068.
StataCorp (2013). Stata Statistical Software: Release 13 (College Station, TX: StataCorp LP).
StataCorp.2013. Stata Statistical Software: Release 13, T. College Station, ed. (StataCorp LP).
Steele, E., and Tucker, A. (2008). Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. Journal of biomedical informatics *41*, 914-926.
Stevens, J.R., and Doerge, R.W. (2005). Combining Affymetrix microarray results. BMC bioinformatics *6*, 57.
Stouffer, S.A., Suchman, E.A., De Vinney, L., Star, S.A., and Williams, R. (1951). Studies in Social Psychology in World War II. Vol. I: The American Soldier: Adjustment during Army Life.
Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Archives of general psychiatry *60*, 1187-1192.
Sutton, A.J., and Abrams, K.R. (2001). Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res *10*, 277-303.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., *et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research *43*, D447-452.

Tan, Y.D., Fornage, M., and Fu, Y.X. (2006). Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes. Genomics *88*, 846-854.

Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D., and Attia, J. (2005). A method for meta-analysis of molecular association studies. Statistics in medicine *24*, 1291-1306.

Thompson, M.D., Hendy, G.N., Percy, M.E., Bichet, D.G., and Cole, D.E. (2014). G protein-coupled receptor mutations and human genetic disease. Methods in molecular biology (Clifton, NJ *1175*, 153-187.

Tremlett, H.L., and Devonshire, V.A. (2006). Does the season or month of birth influence disease progression in multiple sclerosis? Neuroepidemiology *26*, 195-198.

Trzaskowski, B., Latek, D., Yuan, S., Ghoshdastider, U., Debinski, A., and Filipek, S. (2012). Action of molecular switches in GPCRs--theoretical and experimental studies. Current medicinal chemistry *19*, 1090-1109.

Tsai, C.A., Chen, Y.J., and Chen, J.J. (2003). Testing for differentially expressed genes with microarray data. Nucleic Acids Res *31*, e52.

Tsantes, A.E., Nikolopoulos, G.K., Bagos, P.G., Rapti, E., Mantzios, G., Kapsimali, V., and Travlou, A. (2007). Association between the plasminogen activator inhibitor-1 4G/5G polymorphism and venous thrombosis. A meta-analysis. Thrombosis and haemostasis *97*, 907-913.

Tseng, G.C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res *40*, 3785-3799.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

UniProt, C. (2015). UniProt: a hub for protein information. Nucleic acids research *43*, D204-212.

UniProt_Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research *40*, D71-75.

Vinayagam, A., Gibson, T.E., Lee, H.J., Yilmazel, B., Roesel, C., Hu, Y., Kwon, Y., Sharma, A., Liu, Y.Y., Perrimon, N., *et al.* (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proceedings of the National Academy of Sciences of the United States of America *113*, 4976-4981.

Vischer, H.F., Siderius, M., Leurs, R., and Smit, M.J. (2014). Herpesvirus-encoded GPCRs: neglected players in inflammatory and proliferative diseases? Nature reviews Drug discovery *13*, 123-139.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic acids research *41*, W77-83.

Wang, M., and Liu, G. (2015). A simple two-sample Bayesian t-test for hypothesis testing. The American Statistician, 1-20.

Wang, X., Kang, D.D., Shen, K., Song, C., Lu, S., Chang, L.C., Liao, S.G., Huo, Z., Tang, S., Ding, Y., *et al.* (2012a). An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics *28*, 2534-2536.

Wang, Z., Liu, J., Yu, Y., Chen, Y., and Wang, Y. (2012b). Modular pharmacology: the next paradigm in drug discovery. Expert opinion on drug discovery *7*, 667-677.

Watson, J.D. (1987). Molecular Biology of the Gene (Benjamin/Cummings Publishing Company).

Wei, J., and Hemmings, G.P. (2004). Lack of a genetic association between the frizzled-3 gene and schizophrenia in a British population. Neuroscience letters *366*, 336-338.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research *42*, D1001-1006.

Wetzels, R., Raaijmakers, J.G.W., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. Psychonomic Bulletin & Review *16*, 752-760.

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114

Witten, D., and Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. Analysis *17*.
Wong, N.D. (2014). Epidemiological studies of CHD and the evolution of preventive cardiology. Nature reviews Cardiology *11*, 276-289.
Wong, S.K. (2003). G protein selectivity is regulated by multiple intracellular regions of GPCRs. Neuro-Signals *12*, 1-12.
Xi, B., Shen, Y., Reilly, K.H., Wang, X., and Mi, J. (2013). Recapitulation of four hypertension susceptibility genes (CSK, CYP17A1, MTHFR, and FGF5) in East Asians. Metabolism: clinical and experimental *62*, 196-203.
Yang, H., and Churchill, G. (2007). Estimating p-values in small microarray experiments. Bioinformatics *23*, 38-43.
Yang, J., Si, T., Ling, Y., Ruan, Y., Han, Y., Wang, X., Zhang, H., Kong, Q., Li, X., and Liu, C. (2003). Association study of the human FZD3 locus with schizophrenia. Biological psychiatry *54*, 1298-1301.
Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., and Weir, B.S. (2002). Truncated product method for combining P-values. Genet Epidemiol *22*, 170-185.
Zeeberg, B.R., Riss, J., Kane, D.W., Bussey, K.J., Uchio, E., Linehan, W.M., Barrett, J.C., and Weinstein, J.N. (2004). Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC bioinformatics *5*, 80.
Zhang, X., and Eggert, U.S. (2013). Non-traditional roles of G protein-coupled receptors in basic cell biology. Molecular bioSystems *9*, 586-595.
Zhang, Y., Yu, X., Yuan, Y., Ling, Y., Ruan, Y., Si, T., Lu, T., Wu, S., Gong, X., and Zhu, Z. (2004). Positive association of the human frizzled 3 (FZD3) gene haplotype with schizophrenia in Chinese Han population. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics *129*, 16-19.
Zhou, B., Shi, J., and Whittemore, A.S. (2011). Optimal methods for meta-analysis of genome-wide association studies. Genet Epidemiol *35*, 581-591.
Zhou, Q., Rammohan, K., Lin, S., Robinson, N., Li, O., Liu, X., Bai, X.F., Yin, L., Scarberry, B., Du, P., *et al.* (2003). CD24 is a genetic modifier for risk and progression of multiple sclerosis. Proceedings of the National Academy of Sciences of the United States of America *100*, 15041-15046.
Zhou, T.B., Yin, S.S., and Qin, Y.H. (2013). Association of angiotensinogen M235T gene polymorphism with end-stage renal disease risk: a meta-analysis. Molecular biology reports *40*, 765-772.
Zhou, T.B., Yin, S.S., and Qin, Y.H. (2014). Association between angiotensin-converting enzyme insertion/deletion gene polymorphism and end-stage renal disease susceptibility. Journal of the renin-angiotensin-aldosterone system : JRAAS *15*, 22-31.
Zintzaras, E., and Ioannidis, J.P. (2008). Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. Computational biology and chemistry *32*, 38-46.
Μπάγκος, Π. (2015). Βιοπληροφορική (Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών).

Institutional Repository - Library & Information Centre - University of Thessaly
20/04/2024 04:58:49 EEST - 3.135.198.114