



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ
ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

**Ανάπτυξη Λογισμικού για την Διεξαγωγή Μετα-Ανάλυσης
GWAS Δεδομένων**

Γεώργιος Γαλανόπουλος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Παντελεήμων Μπάγκος
Καθηγητής

Λαμία, 2020



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ
ΒΙΟΙΑΤΡΙΚΗ**

**Ανάπτυξη Λογισμικού για την Διεξαγωγή Μετα-Ανάλυσης
GWAS Δεδομένων**

Γεώργιος Γαλανόπουλος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Επιβλέπων
Παντελεήμων Μπάγκος
Καθηγητής**

Λαμία, 2020

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

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 16 /03/2020

Ο Δηλ.

(Υπογραφή)

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

Ανάπτυξη Λογισμικού για την Διεξαγωγή Μετα-Ανάλυσης GWAS Δεδομένων

Γεώργιος Γαλανόπουλος

Τριμελής Επιτροπή:

Παντελεήμων Μπάγκος, Καθηγητής (επιβλέπων)

Γεωργία Μπράλιου, Επίκουρος Καθηγητής

Σωτήριος Τασουλής, Επίκουρος Καθηγητής

Περίληψη

Η μετα-ανάλυση Μελετών Συσχέτισης Ολόκληρου του Γονιδιώματος (Genome Wide Association) έχει ως στόχο την έγκυρη εξαγωγή συμπερασμάτων, πάνω σε ένα συγκεκριμένο αντικείμενο μελέτης και η οποία επιτυγχάνεται από την κοινή επεξεργασία πλήθους ερευνών πάνω στο υπό μελέτη αντικείμενο, μέσω διάφορων στατιστικών μεθόδων. Η διαδικασία αυτή είναι δεδομένα χρονοβόρα και δημιουργεί ανάγκες για ταχύτερη επεξεργασία δεδομένων και εξαγωγή αποτελεσμάτων έτσι ώστε να επιταχύνεται και να διευκολύνεται η ερευνητική διαδικασία.

Η μετα-ανάλυση δεδομένων αυτού του τύπου υλοποιείται από διαφορά υπάρχοντα λογισμικά όπως είναι επί παραδείγματι το GWAMA και το METAL. Τα προγράμματα αυτά έχουν τις δικές τους ιδιαιτερότητες όσον αφορά τους χρόνους εκτέλεσης, καθώς και το πλήθος ή τη μορφή των δεδομένων εξόδου όπως επίσης διαφέρουν και στην μορφή που δέχονται τα ερευνητικά δεδομένα. Γίνεται λοιπόν εύκολα αντιληπτό, πως υπάρχει η ανάγκη υλοποίησης ενός λογισμικού που να συνδυάζει την ταχύτητα εκτέλεσης με την ανάλογη παραγωγή δεδομένων που να εξυπηρετούν άμεσα τον χρήστη δίνοντάς του ποικίλα δεδομένα που χρησιμεύουν στη μετέπειτα έρευνα.

Ο σκοπός της διπλωματικής αυτής εργασίας συνεπώς είναι διττός. Αφενός, είναι η ανάπτυξη ενός λογισμικού που θα είναι ικανό να αυτοματοποιήσει την υπολογιστική διαδικασία της μετα-ανάλυσης με την εξαγωγή πληθώρας δεδομένων που θα εξυπηρετούν και θα απλοποιούν την κατανόηση των αποτελεσμάτων κατά το μέγιστο βαθμό. Αφετέρου, το λογισμικό αυτό έχει ως στόχο να πραγματοποιεί τη διαδικασία σε όσο το δυνατόν μικρότερο χρόνο, διευκολύνοντας έτσι την ερευνητική διαδικασία.

Λέξεις Κλειδιά

Μελέτες Συσχέτισης Ολόκληρου Γονιδιώματος, GWAS Μετα-ανάλυση, Παραλληλοποίηση, Java, Multithreading

Abstract

Genome Wide Association Studies' Meta-Analysis aim to extract valid conclusions by collectively processing all the available studies conducted on the selected query, through certain statistical methods. This process is undoubtedly time-consuming and raises the need for faster data processing in order to acquire the results as quickly as possible, thus expedite the scientific research.

GWAS Meta-Analysis is conducted by several software like GWAMA, METAL etc. These software have their own specific features as far as data processing times are concerned as well as how they choose to input the research data and present the output to the user. It is easily understood, that there is a need to produce a software that combines both fast processing times and a high quality outcome to the researcher in a manner that the variety of the results that are produced is such that the researcher's work is significantly assisted and reduced.

Thus the purpose of this thesis is dual. The first is to develop a software that will be able to automate the computational part of meta-analysis and produce a variety of data that will simplify the understanding of the results in a great way. The second is that the produced software will achieve this in the least possible time thus simplifying the scientific process in a more efficient and time-saving manner.

Keywords

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχάς να ευχαριστήσω τον καθηγητή κ. Παντελή Μπάγκο για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω υπό την επίβλεψη του. Επίσης ευχαριστώ ιδιαίτερα τον κύριο Ιωάννη Ταμπόση για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε καθώς και τους συναδέλφους μου Σοφία Ντέλη και Νίκο Κυριακάκη για την πολύτιμη συνεργασία τους. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Λαμία, Φεβρουάριος 2020

Γεώργιος Γαλιανόπουλος

Περιεχόμενα

Περίληψη	1
Abstract	3
Ευχαριστίες	7
1 Εισαγωγή	13
1.1 Οργάνωση της Εργασίας	13
I Θεωρητικό Μέρος	15
2 Θεωρητικό υπόβαθρο	17
2.1 Η γλώσσα προγραμματισμού JAVA	17
2.1.1 Οι κλάσεις στη JAVA	17
2.1.2 Τα αντικείμενα στη JAVA	18
2.1.3 Το Fork/Join framework στη JAVA	18
2.2 Genome Wide Assosiation Studies (GWAS)	19
2.2.1 Επιστημονικό Υπόβαθρο GWAS	20
2.2.2 Μεθοδολογία για την πραγματοποίηση GWAS	21
2.2.3 Αποτελέσματα GWAS	23
2.2.4 Κλινικές Εφαρμογές GWAS	24
2.2.5 Περιορισμοί των GWAS	24
2.3 Μοντέλα Κληρονομικότητας	25
2.4 Μέτα-Ανάλυση & Συστηματική Ανασκόπηση	26
2.4.1 Η μέτα-ανάλυση στην υγεία	27
2.4.2 Η μεθοδολογία της μέτα-ανάλυσης	28
II Πρακτικό Μέρος	37
3 Περιγραφή Λογισμικού	39
3.1 Δεδομένα εισόδου	40
3.2 Ανάλυση Κλάσεων	41
3.2.1 DataRead.java	41
3.2.2 MetaStats.java	41
3.2.3 ContMetaStats.java	42

3.2.4 PreferenceHandling.java	42
4 Έλεγχος	43
4.1 Μεθοδολογία Ελέγχου	43
4.2 Αναλυτική παρουσίαση ελέγχου	43
4.3 Χρονικές Αποδόσεις	45
III Επίλογος	47
5 Συμπεράσματα	49
5.1 Μελλοντικές Επεκτάσεις	50
Βιβλιογραφία	52
Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια	53
Απόδοση ξενόγλωσσων όρων	55

Κατάλογος Εικόνων

2.1	Κύκλος ζωής ενός Thread	19
2.2	Οι μελέτες GWAS τυπικά προσδιορίζουν κοινές παραλλαγές με μικρά Effect Sizes	21
2.3	Διάγραμμα L' Abbé, με τις τιμές που αντιστοιχούν στις 20 μελέτες να παρουσιάσουν μεγάλη συγκέντρωση, γεγονός που υποδηλώνει την ύπαρξη ομοιογένειας.	30
2.4	Διάγραμμα L' Abbé, με τις τιμές που αντιστοιχούν στις 20 μελέτες να παρουσιάσουν μικρή συγκέντρωση, γεγονός που υποδηλώνει την ύπαρξη ετερογένειας.	31
3.1	Διάγραμμα Ροής του Λογισμικού GWASMetaAnalysis	39
3.2	Δομή αρχείου επιλογών χρήστη	40
3.3	Δομή αρχείου διακριτών δεδομένων εισόδου χρήστη	40
3.4	Δομή αρχείου συνεχών δεδομένων εισόδου χρήστη	41
4.1	Παράδειγμα εκτέλεσης του Stata και τα αποτελέσματα του.	44
4.2	Παράδειγμα εκτέλεσης του λογισμικού μας και τα αποτελέσματα του.	44
4.3	Αρχείο εξόδου του λογισμικού.	45
4.4	Χρόνος εκτέλεσης προγράμματος με 1000 πολυμορφισμούς.	45
4.5	Χρόνος εκτέλεσης προγράμματος με 100.000 πολυμορφισμούς.	46
4.6	Χρόνος εκτέλεσης προγράμματος με 1.000.000 πολυμορφισμούς.	46

Κεφάλαιο 1

Εισαγωγή

Το αντικείμενο μελέτης της παρούσας εργασίας είναι η ανάπτυξη ενός λογισμικού για την αυτοματοποιημένη πραγματοποίηση μέτα-ανάλυσης σε δεδομένα από Genome Wide Association μελέτες ή εν συντομία GWAS. Στην εργασία αυτή τέθηκε το εξής ερώτημα: Είναι δυνατόν να αναπτυχθεί ένα λογισμικό το οποίο να μπορεί να πραγματοποιεί αυτοματοποιημένα μέτα-ανάλυση δεδομένων GWAS, σε χρόνο σημαντικά μικρότερο από τα υπάρχοντα εργαλεία για την περάτωση της παραπάνω διαδικασίας; Το ερώτημα αυτό γεννάται από την ανάγκη των επιστημόνων για επεξεργασία μεγάλου όγκου δεδομένων σε σύντομο χρονικό διάστημα, με σκοπό την πραγματοποίηση μεγαλύτερου πλήθους ερευνών. Αρχικά, η πρώτη προτεραιότητα για την εκκίνηση της συγγραφής του εν λόγω προγράμματος, ήταν η επιλογή της γλώσσας προγραμματισμού που θα χρησιμοποιηθεί. Η επιτακτική ανάγκη για τη χρήση μιας γλώσσας η οποία να συνδυάζει υψηλή απόδοση σε συνδυασμό με πληθώρα πακέτων για την διευκόλυνση της συγγραφής του λογισμικού, οδήγησε στην επιλογή της γλώσσας προγραμματισμού Java.

Εν συνεχεία, ερευνήθηκε το περιεχόμενο του λογισμικού αυτού, οι επιλογές δηλαδή που θα έχει ο χρήστης όσον αφορά τις στατιστικές συναρτήσεις, τα μαθηματικά μοντέλα και τα δεδομένα εισόδου που θα δίδονται από εκείνον. Ο περιορισμένος όγκος βιβλιογραφίας σχετικά με τη μέτα-ανάλυση των GWAS δεδομένων κυρίως όσον αφορά το κομμάτι της μεθοδολογίας που χρησιμοποιείται για την πραγματοποίηση της, αποτέλεσε αρχικά τροχοπέδη στη συγγραφή της παρούσας εργασίας αλλά με την πολύτιμη βοήθεια του επιβλέποντα καθηγητή κ. Μπάγκου καθώς και λαμβάνοντας υπόψη τις ήδη υπάρχουσες υλοποιήσεις από υπολογιστικά πακέτα όπως το METAL το MetABEL και το GWAMA καταλήξαμε στην υπάρχουσα δομή η οποία θα αναλυθεί παρακάτω.

Τέλος, ένα μείζον θέμα που μελετάται και επιλύεται από την παρούσα εργασία είναι ο τρόπος με τον οποίο η διαδικασία της μετανάλυσης θα γίνει με τον ταχύτερο δυνατό τρόπο. Για το σκοπό αυτό χρησιμοποιήθηκε η παραλληλοποίηση και πιο συγκεκριμένα `tofork/join framework` που παρέχεται από την JAVA.

1.1 Οργάνωση της Εργασίας

Η εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια: Στο Κεφάλαιο 2 δίνεται το θεωρητικό υπόβαθρο των βασικών τεχνολογιών που σχετίζονται με τη διπλωματική αυτή. Αρχικά υπάρχει μια σύντομη περιγραφή της γλώσσας προγραμματισμού JAVA και μια ανάλυση της

εφαρμογής της στο παρόν λογισμικό. Επιπλέον στο ίδιο κεφάλαιο αναλύεται η έννοια των GWAS και η θεωρία πίσω απ' αυτά καθώς επίσης αναλύεται και η έννοια της μέτα-ανάλυσης και παρουσιάζεται ενδελεχώς η μεθοδολογία με την οποία υλοποιείται. Στο Κεφάλαιο 3 παρουσιάζεται η ανάλυση και η σχεδίαση του λογισμικού, δηλαδή η περιγραφή της δομής του, των συναρτήσεων και των βασικών αλγορίθμων που έχουν χρησιμοποιηθεί. Στο Κεφάλαιο 4 παρουσιάζεται ο έλεγχος καλής λειτουργίας του συστήματος με βάση ένα συγκεκριμένο σενάριο χρήσης. Τέλος στο Κεφάλαιο 5 δίνεται η συνεισφορά αυτής της διπλωματικής εργασίας, καθώς και μελλοντικές επεκτάσεις.

Μέρος I

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται αναλυτικά οι τρεις βασικοί θεωρητικοί πυλώνες πάνω στους οποίους έχει στηριχθεί η ανάπτυξη του παρόντος λογισμικού, δηλαδή η γλώσσα προγραμματισμού JAVA, η έννοια της μέτα-ανάλυσης και τα GWAS.

2.1 Η γλώσσα προγραμματισμού JAVA

Η JAVA είναι μία γλώσσα προγραμματισμού γενικού σκοπού και συγκαταλέγεται στις αντικειμενοστρεφείς γλώσσες ενώ είναι βασισμένη σε κλάσεις. Είναι σχεδιασμένη με τέτοιο τρόπο ώστε να έχει όσες δυνατόν λιγότερες εξαρτήσεις. Ένα ακόμη σημαντικό χαρακτηριστικό της γλώσσας αυτής είναι πως επιτρέπει στους προγραμματιστές εφόσον γράψουν μια εφαρμογή και κάνουν compile σε κάποιο μηχάνημα, να έχουν τη δυνατότητα να μπορούν να την “τρέξουν” σε οποιοδήποτε άλλο σύστημα υποστηρίζει τη JAVA χωρίς να υπάρχει ανάγκη να γίνει compile εκ νέου[1]. Τα βασικά χαρακτηριστικά της γλώσσας που χρησιμοποιήθηκαν κατά την ανάπτυξη του λογισμικού ήταν τα εξής:

- Αντικείμενα.
- Κλάσεις.
- Το Fork/Join Framework.

2.1.1 Οι κλάσεις στη JAVA

Μια κλάση είναι ένα σύνολο δηλώσεων που αφορούν στην περιγραφή μιας συγκεκριμένης κατηγορίας αντικειμένων. Πιο πρακτικά, μέσα σε μία κλάση αποσαφηνίζονται τα χαρακτηριστικά γνωρίσματα και οι μέθοδοι όλων των αντικειμένων που δύνανται να παραχθούν από τη συγκεκριμένη κλάση.

Ένα χαρακτηριστικό παράδειγμα κλάσης θα μπορούσε να είναι ένα σπίτι:

1. class House{
2. String owner;
3. String address;
4. int postcode;
5. float sqmetres;
6. }

Όπως γίνεται εύκολα αντιληπτό και στο άνωθεν παράδειγμα σε μία κλάση Σπίτι(House) ορίζει όλες τις παραμέτρους που χρειάζονται για να προσδιορίσουμε ένα αντικείμενο της συγκεκριμένης κλάσης. Δηλαδή:

- Τον ιδιοκτήτη.
- Τη διεύθυνση.
- Τον ταχυδρομικό κώδικα.
- Τα τετραγωνικά μέτρα.

2.1.2 Τα αντικείμενα στη JAVA

Τα αντικείμενα είναι η βασική μονάδα του λεγόμενου αντικειμενοστραφούς προγραμματισμού και αντιπροσωπεύουν την "κατάσταση" ή την "συμπεριφορά" μίας οντότητας μίας κλάσης. Ένα αντικείμενο ορίζεται ως ένα σύνολο από πεδία (fields) (ή μεταβλητές υπόστασης (instance variables) ή ιδιότητες (properties) και μεθόδους (methods)).[1] Και οι δύο κατηγορίες ονομάζονται μέλη (members) του αντικειμένου. Τα πεδία περιέχουν δεδομένα σχετικά με την κατάσταση του αντικειμένου ενώ οι μέθοδοι περιέχουν κώδικα που επιτρέπει την πρόσβαση και την αλλαγή των ιδιοτήτων ενός αντικειμένου. Τα πεδία και οι μέθοδοι ενός συγκεκριμένου αντικειμένου ονομάζονται πεδία και ιδιότητες της υπόστασης (ινστανς) του.

2.1.3 Το Fork/Join framework στη JAVA

Η Θαα είναι μια πολυνηματική γλώσσα προγραμματισμού. Ένα πρόγραμμα με πολλά νημάτα(threads) περιέχει δύο ή περισσότερα τμήματα που μπορούν να τρέχουν ταυτόχρονα και κάθε ένα από αυτά μπορεί να χειριστεί μια διαφορετική εργασία ταυτόχρονα, κάνοντας τη βέλτιστη χρήση των διαθέσιμων πόρων ειδικά όταν ο υπολογιστής έχει πολλαπλές CPUs.

Εξ ορισμού, το multitasking είναι όταν πολλές διαδικασίες μοιράζονται κοινούς πόρους επεξεργασίας, όπως μια CPU. Το Multi-threading επεκτείνει την ιδέα του multitasking σε εφαρμογές όπου μπορούμε να υποδιαιρέσουμε συγκεκριμένες λειτουργίες από μια ενιαία εφαρμογή σε μεμονωμένα νήματα.

Κάθε ένα από τα νήματα μπορεί να τρέξει παράλληλα. Το λειτουργικό σύστημα διαιρεί τον χρόνο επεξεργασίας όχι μόνο μεταξύ διαφορετικών εφαρμογών, αλλά και μεταξύ κάθε νήματος μέσα σε μια εφαρμογή.

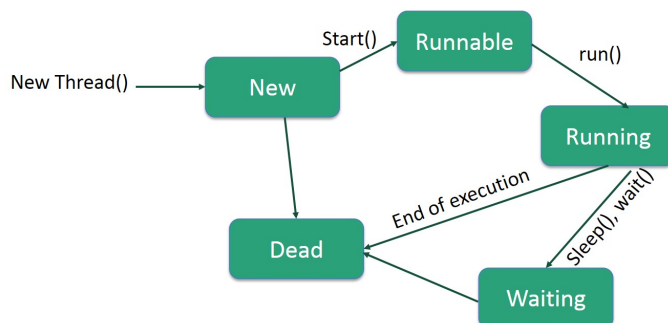
Ένα νήμα περνάει από διάφορα στάδια του κύκλου ζωής του. Για παράδειγμα, ένα νήμα γεννιέται, ξεκινά, τρέχει και στη συνέχεια πεθαίνει. Το παρακάτω διάγραμμα δείχνει τον πλήρη κύκλο ζωής ενός νήματος.

Ακολουθούν τα στάδια του κύκλου ζωής ενός νήματος:

- **New**- Ένα νέο νήμα ξεκινά τον κύκλο ζωής του στη νέα κατάσταση. Παραμένει σε αυτήν την κατάσταση μέχρι να ξεκινήσει το πρόγραμμα το νήμα.
- **Runnable**: Μετά την εκκίνηση ενός "νεογέννητου" νήματος, το νήμα γίνεται εκτελέσιμο. Ένα νήμα σε αυτή την κατάσταση θεωρείται ότι εκτελεί την προκαθορισμένη εργασία του.

- **Waiting:** Ενίοτε, ένα νήμα μεταβαίνει στην κατάσταση αναμονής όσο περιμένει ένα άλλο νήμα να εκτελέσει μια εργασία. Ένα νήμα μεταβαίνει πίσω στην τρέχουσα κατάσταση μόνο όταν ένα άλλο νήμα σηματοδοτεί στο νήμα που περιμένει να συνεχιστεί η εκτέλεση.
- **Timed Waiting:** Ένα τρέχον νήμα μπορεί να εισέλθει στη χρονική κατάσταση αναμονής για συγκεκριμένο χρονικό διάστημα. Ένα νήμα σε αυτή την κατάσταση μεταβαίνει πίσω στην τρέχουσα κατάσταση όταν αυτό το χρονικό διάστημα λήγει ή όταν συμβαίνει το γεγονός που περιμένει.
- **Terminated(Dead):** Ένα τρέχον νήμα εισέρχεται στην τερματισμένη κατάσταση όταν ολοκληρώνει την εργασία του ή τελειώνει με άλλο τρόπο.

[2] [3]



Εικόνα 2.1: Κύκλος ζωής ενός Thread

Κάθε νήμα στη Θαα έχει μία προτεραιότητα που βοηθά το λειτουργικό σύστημα να καθορίσει τη σειρά με την οποία θα προγραμματιστούν τα νήματα. Οι προτεραιότητες των νημάτων στη Θαα κυμαίνονται στην περιοχή μεταξύ $MIN_{PRIORITY}$ (μια σταθερά 1) και $MAX_{PRIORITY}$ (μια σταθερά 10). Από προεπιλογή, κάθε νήμα έχει προτεραιότητα $NORM_{PRIORITY}$ (μια σταθερά 5).

Τα νήματα με υψηλότερη προτεραιότητα είναι πιο σημαντικά σε ένα πρόγραμμα και θα πρέπει να διατεθεί επεξεργαστικός χρόνος πριν από τα νήματα χαμηλής προτεραιότητας. Ωστόσο, η προτεραιότητα των νημάτων δεν μπορεί να εγγυηθεί τη σειρά με την οποία τα νήματα θα εκτελεστούν διοτί αυτό καθορίζεται και σε μεγάλο βαθμό από την πλατφόρμα.

2.2 Genome Wide Assosiation Studies (GWAS)

Στη γενετική, μια μελέτη συσχέτισης ολόκληρου του γονιδιώματος (μελέτη GWAS ή GWAS), επίσης γνωστή ως μελέτη πλήρους γονιδιώματος (μελέτη WGA ή WGAS), είναι μια παρατηρητική μελέτη ενός γενωμικού συνόλου γενετικών παραλλαγών σε διαφορετικά άτομα για να διαπιστωθεί εάν οποιαδήποτε παραλλαγή συνδέεται με ένα συγκεκριμένο χαρακτηριστικό. Τα GWAS επικεντρώνονται συνήθως σε συσχετισμούς μεταξύ πολυμορφισμών ενός νουκλεοτιδίου (SNP) και χαρακτηριστικών όπως οι κύριες ανθρώπινες ασθένειες, αλλά μπορούν εξίσου να εφαρμοστούν σε οποιαδήποτε άλλη γενετική παραλλαγή και σε οποιονδήποτε άλλο οργανισμό.

Όταν εφαρμόζεται σε ανθρώπινα δεδομένα, οι μελέτες GWAS συγκρίνουν το DNA των συμμετεχόντων που έχουν διαφορετικούς φαινότυπους για ένα συγκεκριμένο γνώρισμα ή ασθένεια. Αυτοί οι συμμετέχοντες μπορεί να είναι άτομα με νόσο (cases) και παρόμοια άτομα χωρίς τη νόσο (controls), ή μπορεί να είναι άτομα με διαφορετικούς φαινότυπους για ένα συγκεκριμένο χαρακτηριστικό, για παράδειγμα την αρτηριακή πίεση. Αυτή η προσέγγιση είναι γνωστή ως φαινότυπος-πρώτα, στην οποία οι συμμετέχοντες ταξινομούνται πρώτα από την κλινική τους εκδήλωση, σε αντίθεση με τον γονότυπο-πρώτα. Κάθε άτομο δίνει ένα δείγμα DNA, από το οποίο διαβάζονται εκατομμύρια γενετικών παραλλαγών χρησιμοποιώντας συστοιχίες SNP. Εάν ένας τύπος της παραλλαγής (ένα αλληλόμορφο) είναι πιο συχνός σε άτομα με τη νόσο, η παραλλαγή λέγεται ότι σχετίζεται με την ασθένεια. Τα συσχετιζόμενα SNP θεωρούνται στη συνέχεια ότι σηματοδοτούν μια περιοχή του ανθρώπινου γονιδιώματος που μπορεί να επηρεάσει τον κίνδυνο ασθένειας.

Οι μελέτες GWAS διερευνούν ολόκληρο το γονιδίωμα, σε αντίθεση με μεθόδους που δοκιμάζουν συγκεκριμένα έναν μικρό αριθμό προκαθορισμένων γενετικών περιοχών. Ως εκ τούτου, το GWAS είναι μια μη μεροληπτική προς τους υποψήφιους προσέγγιση, σε αντίθεση με τις μελέτες που βασίζονται σε συγκεκριμένα γονίδια. Οι μελέτες GWAS εντοπίζουν SNPs και άλλες παραλλαγές στο DNA που σχετίζονται με μια ασθένεια, αλλά δεν μπορούν από μόνες τους να καθορίσουν ποια γονίδια είναι αιτιώδη.

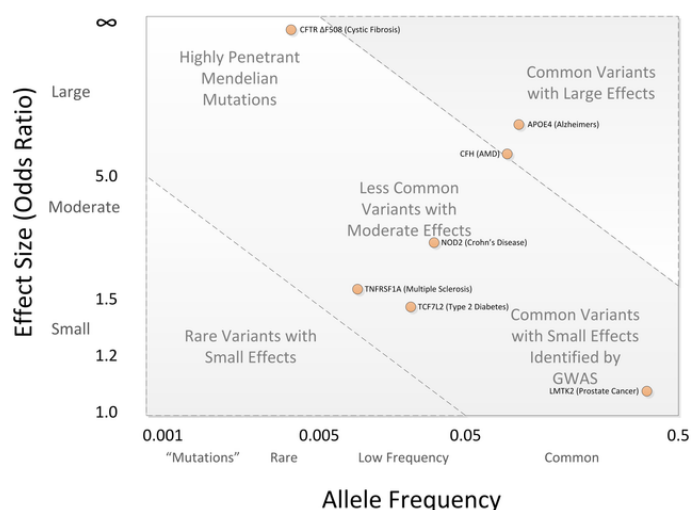
Το πρώτο επιτυχημένο GWAS δημοσιεύτηκε το 2002 και μελέτησε το έμφραγμα του μυοκαρδίου. Αυτός ο σχεδιασμός μελέτης στη συνέχεια εφαρμόστηκε στη μελέτη ορόσημο GWAS 2005 που ερεύνησε ασθενείς με εκφυλισμό της ωχράς κηλίδας που σχετίζεται με την ηλικία και βρήκε δύο SNPs με σημαντικά αλλοιωμένη συχνότητα αλληλόμορφων σε σύγκριση με τους υγιείς μάρτυρες (controls). Από το 2017, πάνω από 3.000 ανθρώπινες μελέτες GWAS έχουν εξετάσει πάνω από 1.800 ασθένειες και γνωρίσματα και έχουν βρεθεί χιλιάδες συσχετίσεις SNP. Σε γενικές γραμμές, οι συσχετίσεις αυτές είναι πολύ αδύναμες και μπορεί ακόμη και να μην έχουν νόημα, εκτός από τις σπάνιες γενετικές ασθένειες.[4]

2.2.1 Επιστημονικό Υπόβαθρο GWAS

Οποιαδήποτε δύο ανθρώπινα γονιδιώματα διαφέρουν με εκατομμύρια διαφορετικούς τρόπους. Υπάρχουν μικρές παραλλαγές στα μεμονωμένα νουκλεοτίδια των γονιδιωμάτων (SNP's) καθώς και πολλές μεγαλύτερες παραλλαγές, όπως διαγραφές, εισαγωγές και παραλλαγές αριθμού αντιγράφων. Οποιοδήποτε από αυτά μπορεί να προκαλέσει αλλοιώσεις στα χαρακτηριστικά ενός ατόμου ή στο φαινότυπο του, το οποίο μπορεί να επηρεάσει οτιδήποτε από τον κίνδυνο εμφάνισης ασθένειας έως τις φυσικές ιδιότητες όπως το ύψος. Περίπου το έτος 2000, πριν από την εισαγωγή των μελετών GWAS, η βασική μέθοδος διερεύνησης ήταν μέσω κληρονομικών μελετών της γενετικής σύνδεσης στις οικογένειες. Αυτή η προσέγγιση αποδείχθηκε ιδιαίτερα χρήσιμη για τις διαταραχές ενός γονιδίου. Ωστόσο, για κοινές και πολύπλοκες ασθένειες τα αποτελέσματα των γενετικών μελετών έδειξαν ότι είναι δύσκολο να αναπαραχθούν. Μια εναλλακτική πρόταση στις μελέτες σύνδεσης ήταν η γενετική μελέτη σύνδεσης. Αυτός ο τύπος μελέτης ρωτά αν το αλληλόμορφο μιας γενετικής παραλλαγής βρίσκεται συχνότερα από το αναμενόμενο σε άτομα με τον φαινότυπο ενδιαφέροντος (π.χ. με την ασθένεια που μελετάται). Οι πρώτοι υπολογισμοί σχετικά με τη στατιστική ισχύ έδειξαν

ότι αυτή η προσέγγιση θα μπορούσε να είναι καλύτερη από τις μελέτες σύνδεσης κατά την ανίχνευση ασθενών γενετικών επιδράσεων.

Εκτός από το εννοιολογικό πλαίσιο αρκετοί πρόσθετοι παράγοντες πυροδότησαν τις μελέτες GWAS. Το ένα ήταν η εμφάνιση των βιοτραπεζών, οι οποίες αποτελούν αποθέματα ανθρώπινου γενετικού υλικού, που μείωσαν σημαντικά το κόστος και τη δυσκολία συγκέντρωσης επαρκούς αριθμού βιολογικών δειγμάτων για μελέτη. Ένα άλλο ήταν το Διεθνές Πρόγραμμα HarMap, το οποίο, από το 2003, αναγνώρισε την πλειονότητα των κοινών SNP που αναλύθηκαν σε μια μελέτη GWAS. Η δομή haploblock που προσδιορίστηκε από το έργο HarMap επέτρεψε επίσης την εστίαση στο υποσύνολο των SNP που θα περιγράφουν το μεγαλύτερο μέρος της παραλλαγής. Επίσης, η ανάπτυξη των μεθόδων παραγωγής γονότυπου από όλα αυτά τα SNPs χρησιμοποιώντας συστοιχίες γονότυπου ήταν μια σημαντική προϋπόθεση.[4]



Εικόνα 2.2: Οι μελέτες GWAS τυπικά προσδιορίζουν κοινές παραληλαγές με μικρά Effect Sizes

2.2.2 Μεθοδολογία για την πραγματοποίηση GWAS

Η πιο συνηθισμένη προσέγγιση των μελετών GWAS είναι το μοντέλο case-control, το οποίο συγκρίνει δύο μεγάλες ομάδες ατόμων, μία υγιή ομάδα (control) και μία ομάδα ασθενών που πάσχουν από ασθένεια (cases). Όλα τα άτομα σε κάθε ομάδα γονοτυπούνται για την πλειονότητα των κοινότυπων γνωστών SNP. Ο ακριβής αριθμός των SNPs εξαρτάται από την τεχνολογία που γονοτυπούνται, αλλά είναι τυπικά ένα εκατομμύριο ή περισσότερα. Για κάθε ένα από αυτά τα SNPs διερευνάται εάν η συχνότητα αλληλόμορφων μεταβάλλεται σημαντικά μεταξύ των cases και της ομάδας control. Σε αυτά τα μοντέλα, η θεμελιώδης μονάδα για την αναφορά μεγεθών αποτελεσμάτων είναι το odds ratio.

Το odds ratio είναι ο λόγος δύο σχετικών πιθανοτήτων, οι οποίες στο πλαίσιο των μελετών GWAS είναι οι πιθανότητες τα άτομα case να έχουν ένα συγκεκριμένο αλληλόμορφο και τις πιθανότητες τα άτομα control να μην έχουν το ίδιο αλληλόμορφο. Όταν η συχνότητα αλληλόμορφων στην ομάδα cases είναι πολύ υψηλότερη από ό, τι στην ομάδα control, το odds ratio είναι μεγαλύτερο από 1 και αντίστροφα για τη χαμηλότερη συχνότητα αλληλόμορφων. Επιπλέον, το P-value για τη στατιστική σημαντικότητα του odds ratio υπολογίζεται συνήθως

χρησιμοποιώντας ένα απλό Chi-squared test. Η εύρεση odds ratio που διαφέρουν σημαντικά από το 1 είναι ο στόχος της μελέτης GWAS επειδή αυτό δείχνει ότι ένα SNP σχετίζεται με ασθένεια.

Υπάρχουν πολλές παραλλαγές σε αυτήν την προσέγγιση case-control. Μια συνήθης εναλλακτική στις case-control GWAS μελέτες, είναι η ποσοτική ανάλυση φαινοτυπικών δεδομένων, όπως π.χ. το ύψος, ή κάποιοι βιοδείκτες ή ακόμη και η γενετική έκφραση. Ομοίως, μπορούν να χρησιμοποιηθούν εναλλακτικές στατιστικές που έχουν σχεδιαστεί για κυριαρχικά ή υπολειπόμενα μοτίβα διείσδυσης. Οι υπολογισμοί γίνονται συνήθως χρησιμοποιώντας λογισμικό βιοπληροφορικής, όπως το SNPTTEST[5] και το PLINK[6], το οποίο υποστηρίζει πολλές από αυτές τις εναλλακτικές στατιστικές. Παλαιότερα το GWAS επικεντρώθηκε στην επίδραση των μεμονωμένων SNP. Ωστόσο, οι εμπειρικές αποδείξεις δείχνουν ότι πολύπλοκες αλληλεπιδράσεις μεταξύ δύο ή περισσότερων SNP, της λεγόμενης και ως Επίστασης, μπορεί να συμβάλλουν σε πολύπλοκες ασθένειες. Επιπλέον, οι ερευνητές προσπαθούν να ενσωματώσουν τα δεδομένα GWAS με άλλα βιολογικά δεδομένα όπως το δίκτυο αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης για να εξαγάγουν περισσότερα πληροφοριακά αποτελέσματα.

Ένα βασικό βήμα στην πλειοψηφία των μελετών GWAS είναι ο καταλογισμός των γονότυπων σε SNPs όχι στο τσιπ γονότυπου που χρησιμοποιήθηκε στη μελέτη. Αυτή η διαδικασία αυξάνει σημαντικά τον αριθμό των SNP που μπορούν να δοκιμαστούν για συσχέτιση, αυξάνουν τη δύναμη της μελέτης και διευκολύνουν μετα-ανάλυση του GWAS σε διακριτές κοορτές. Ο καταλογισμός του γονότυπου πραγματοποιείται με στατιστικές μεθόδους που συνδυάζουν τα δεδομένα GWAS μαζί με μια ομάδα αναφοράς απλοτύπων. Αυτές οι μέθοδοι εκμεταλλεύονται την κατανομή των απλοτύπων μεταξύ ατόμων σε σύντομα τμήματα αλληλουχίας για να καταλογίζουν αλλήλια. Τα υπάρχοντα πακέτα λογισμικού για τον καταλογισμό του γονότυπου περιλαμβάνουν τα IMPUTE2, Minimac, Beagle και MaCH.

Εκτός από τον υπολογισμό της συσχέτισης, είναι συνήθης η συνεκτίμηση κάθε μεταβλητής που θα μπορούσε να προκαλέσει σύγχυση στα αποτελέσματα. Το φύλο και η ηλικία είναι συνηθισμένα παραδείγματα συγχυστικών μεταβλητών. Επιπλέον, είναι επίσης γνωστό ότι πολλές γενετικές παραλλαγές συνδέονται με τους γεωγραφικούς και ιστορικούς πληθυσμούς στους οποίους προέκυψαν οι μεταλλάξεις. Λόγω αυτής της συσχέτισης, οι μελέτες πρέπει να λαμβάνουν υπόψη το γεωγραφικό και εθνοτικό υπόβαθρο των συμμετεχόντων ελέγχοντας τη λεγόμενη πληθυσμιακή διαστρωμάτωση. Αν δεν το κάνουν, αυτές οι μελέτες μπορούν να οδηγήσουν σε ψευδώς θετικά αποτελέσματα.

Αφού τα Odds-Ratio και τα P-Values έχουν υπολογιστεί για όλα τα SNPs, μια κοινή προσέγγιση είναι να δημιουργηθεί ένα Manhattan Plot. Στο πλαίσιο των μελετών GWAS, αυτή η γραφική παράσταση δείχνει τον αρνητικό λογάριθμο της τιμής P ως συνάρτηση της γονιδωματικής θέσης. Έτσι, τα SNPs με την πιο σημαντική συσχέτιση ξεχωρίζουν στο γράφημα συνήθως ως στοίβες σημείων λόγω της δομής του haploblock. Επιπλέον είναι σημαντικό να διορθωθεί το κατώφλι του P-Value για τη στατιστική σημαντικότητα για πολλαπλά θέματα δοκιμών. Το ακριβές κατώφλι ποικίλλει ανά μελέτη, αλλά το συμβατικό όριο είναι 5×10^{-8} για να είναι στατιστικά σημαντικό για εκατοντάδες χιλιάδες έως εκατομμύρια δοκιμασμένων SNPs. Οι μελέτες GWAS τυπικά εκτελούν την πρώτη ανάλυση σε μια κοορτή ανίχνευσης, ακολουθούμενη από επικύρωση των σημαντικότερων SNPs σε μια ανεξάρτητη κοορτή επικύρωσης.[7]

2.2.3 Αποτελέσματα GWAS

Έχουν γίνει προσπάθειες για τη δημιουργία ολοκληρωμένων καταλόγων SNP που έχουν εντοπιστεί από τις μελέτες GWAS. Από το 2009, τα SNP που σχετίζονται με ασθένειες αριθμούνται σε χιλιάδες.

Η πρώτη μελέτη GWAS, που διεξήχθη το 2005, συνέκρινε 96 ασθενείς με εκφύλιση της ωχράς κηλίδας (ARMD) με 50 υγιείς μάρτυρες. Εντοπίστηκαν δύο SNPs με σημαντικά αλλοιωμένη συχνότητα αλληλόμορφων μεταξύ των δύο ομάδων. Αυτά τα SNPs εντοπίστηκαν στο γονίδιο που κωδικοποιεί τον παράγοντα συμπληρώματος H, το οποίο ήταν ένα απροσδόκητο εύρημα στην έρευνα του ARMD. Τα ευρήματα από αυτές τις πρώτες μελέτες GWA στη συνέχεια οδήγησαν σε περαιτέρω λειτουργική έρευνα προς το θεραπευτικό χειρισμό του συστήματος συμπληρώματος στο ARMD. Μια άλλη σημαντική αναφορά στην ιστορία των μελετών GWAS ήταν η μελέτη WTCCC, η μεγαλύτερη μελέτη GWAS που διεξήχθη ποτέ κατά τη στιγμή της δημοσίευσής της το 2007. Το WTCCC περιελάμβανε 14.000 περιπτώσεις επτά κοινών ασθενειών (περίπου 2.000 άτομα για καθμία από τις στεφανιαίες καρδιακές παθήσεις, τον διαβήτη τύπου 1, τον διαβήτη τύπου 2, τη ρευματοειδή αρθρίτιδα, τη νόσο του Crohn, τη διπολική διαταραχή και την υπέρταση) και 3.000 κοινό έλεγχο. Η μελέτη αυτή ήταν επιτυχής στην αποκάλυψη πολλών νέων γονιδίων που αποτελούν τη βάση αυτών των ασθενειών.

Από αυτές τις πρώτες μελέτες ορόσημο GWAS, υπήρξαν δύο γενικές τάσεις. Η μία έχει να κάνει με όλο και μεγαλύτερα μεγέθη δειγμάτων. Το 2018, αρκετές μελέτες GWAS φθάνουν σε ένα συνολικό μέγεθος δείγματος άνω του 1 εκατομμυρίου συμμετεχόντων, συμπεριλαμβανομένων 1,1 εκατομμυρίων σε μια γενική μελέτη γονιδιώματος σχετικά με το μορφωτικό επίπεδο και μια μελέτη της αϋπνίας που περιέχει 1,3 εκατομμύρια άτομα. Ο λόγος είναι η κλίση προς την αξιόπιστη ανίχνευση των κινδύνων SNP που έχουν μικρότερες αναλογίες πιθανότητας και χαμηλότερη συχνότητα αλληλόμορφων. Μια άλλη τάση ήταν η χρήση πιο στενά καθορισμένων φαινοτύπων, όπως τα λιπίδια του αίματος, η προϊνσουλίνη ή παρόμοιοι βιοδείκτες. Αυτοί ονομάζονται ενδιάμεσοι φαινότυποι και οι αναλύσεις τους μπορεί να έχουν αξία στη λειτουργική έρευνα σε βιοδείκτες. Μια παραλλαγή του GWAS χρησιμοποιεί συμμετέχοντες που είναι συγγενείς πρώτου βαθμού ατόμων με νόσο. Αυτός ο τύπος μελέτης ονομάστηκε μελέτη γονιδιακής σύνδεσης με μεσολάβηση (GWAX).

Ένα κεντρικό σημείο συζήτησης σχετικά με τις μελέτες GWAS ήταν ότι οι περισσότερες από τις παραλλαγές SNP που εντοπίστηκαν από τις μελέτες GWAS σχετίζονται με ένα μικρό ποσοστό αυξημένου κινδύνου για τη νόσο και έχουν μικρή μόνο προγνωστική αξία. Το μέσο Odds Ratio είναι 1,33 ανά risk-SNP, με λίγα μόνο Odds Ratio πάνω από 3,0. Αυτά τα μεγέθη θεωρούνται μικρά επειδή δεν διακτιολογούν μεγάλο μέρος της κληρονομικής παραλλαγής. Αυτή η κληρονομική παραλλαγή είναι γνωστή από μελέτες κληρονομικότητας που βασίζονται σε μονοζυγωτικά δίδυμα. Για παράδειγμα, είναι γνωστό ότι το 80-90% της διακύμανσης στο ύψος μπορεί να εξηγηθεί από κληρονομικές διαφορές, αλλά οι μελέτες GWAS αντιπροσωπεύουν μόνο μια μειονότητα αυτής της διακύμανσης.[7]

2.2.4 Κλινικές Εφαρμογές GWAS

Μια πρόκληση για τη μελλοντική επιτυχία μιας μελέτης GWAS είναι να εφαρμόσει τα ευρήματα με τρόπο που θα επιταχύνει την ανάπτυξη φαρμάκων και διαγνωστικών μεθόδων, συμπεριλαμβανομένης της καλύτερης ενσωμάτωσης των γενετικών μελετών στη διαδικασία ανάπτυξης φαρμάκων και της εστίασης στο ρόλο της γενετικής ποικιλίας στη διατήρηση της υγείας ως πρότυπο σχεδιάζοντας νέα φάρμακα και διαγνωστικές μεθόδους. Πολλές μελέτες έχουν εξετάσει τη χρήση σημάτων risk-SNP ως μέσο άμεσης βελτίωσης της ακρίβειας της πρόγνωσης. Μερικοί έχουν διαπιστώσει ότι η ακρίβεια της πρόγνωσης βελτιώνεται, ενώ άλλοι αναφέρουν ελάχιστα οφέλη από αυτή τη χρήση. Γενικά, ένα πρόβλημα με αυτή την άμεση προσέγγιση είναι τα μικρά μεγέθη των παρατηρούμενων επιδράσεων. Μία μικρή επίδραση τελικά μπορεί να μεταφραστεί σε έναν κακό διαχωρισμό των ατόμων cases και των ατόμων controls και επομένως μόνο μια μικρή βελτίωση της ακρίβειας της πρόγνωσης. Μια εναλλακτική εφαρμογή είναι επομένως η δυνατότητα των μελετών GWAS να διασαφηνίσουν την παθοφυσιολογία.[7]

2.2.5 Περιορισμοί των GWAS

Οι μελέτες GWAS έχουν διάφορα θέματα και περιορισμούς που μπορούν να επιλυθούν μέσω κατάλληλου ελέγχου ποιότητας και εγκατάστασης μελέτης. Η έλλειψη σαφώς καθορισμένων ομάδων περιπτώσεων και μαρτύρων, το ανεπαρκές μέγεθος δείγματος, ο έλεγχος για πολλαπλές δοκιμές και ο έλεγχος για τη στρωματοποίηση του πληθυσμού είναι συνήθη προβλήματα. Ιδιαίτερα το στατιστικό ζήτημα των πολλαπλών δοκιμών όπου έχει σημειωθεί ότι "η προσέγγιση GWAS μπορεί να είναι προβληματική επειδή ο τεράστιος αριθμός των στατιστικών δοκιμών που εκτελούνται παρουσιάζει μια άνευ προηγουμένου δυνατότητα για ψευδώς θετικά αποτελέσματα". Η αγνόηση αυτών των διορθώσιμων ζητημάτων έχει αναφερθεί ως συμβολή σε μια γενική αίσθηση προβλημάτων με τη μεθοδολογία GWAS. Εκτός από τα εύκολα διορθώσιμα προβλήματα όπως αυτά, έχουν προκύψει κάποια πιο λεπτά αλλά σημαντικά θέματα. Μια περίοπτη μελέτη GWAS που διερεύνησε άτομα με πολύ μεγάλη διάρκεια ζωής για να εντοπίσει τα SNP που σχετίζονται με τη μακροζωία είναι ένα παράδειγμα αυτού. Η δημοσίευση ελέγχθηκε ενδελεχώς εξαιτίας μιας ανισορροπίας μεταξύ του τύπου της σειράς γονότυπων στα άτομα cases και στα άτομα control, η οποία προκάλεσε ψευδείς επισημάνσεις πολλών SNP που σχετίζονταν με τη μακροζωία. Στη συνέχεια η μελέτη αποσύρθηκε, αλλά αργότερα δημοσιεύθηκε τροποποιημένο χειρόγραφο.

Εκτός από αυτά τα θέματα που μπορούν να αποφευχθούν, οι μελέτες GWAS έχουν προσελκύσει πιο θεμελιώδη κριτική, κυρίως λόγω της υπόθεσής τους ότι η κοινή γενετική ποικιλία παίζει σημαντικό ρόλο στην εξήγηση της κληρονομικής παραλλαγής της κοινής ασθένειας. Πράγματι, έχει εκτιμηθεί ότι για τις περισσότερες συνθήκες η κληρονομικότητα που αποδίδεται στο SNP είναι 0,05. Αυτή η πτυχή των μελετών GWAS έχει προσελκύσει την κριτική ότι, αν και δεν ήταν γνωστό μελλοντικά, οι μελέτες GWAS τελικά δεν άξιζαν τις δαπάνες. Οι μελέτες GWAS αντιμετωπίζουν επίσης την κριτική ότι η ευρεία ποικιλία των μεμονωμένων αποκρίσεων ή των αντισταθμιστικών μηχανισμών σε μια κατάσταση ασθένειας ακυρώνει και καλύπτει πιθανά γονίδια ή αιτιακές παραλλαγές που συνδέονται με τη νόσο. Επιπρόσθετα, οι μελέτες GWAS εντοπίζουν υποψήφιες μεταβλητές κινδύνου για τον πλη-

θυσμό από τον οποίο πραγματοποιείται η ανάλυσή τους και με τις περισσότερες μελέτες GWAS που προέρχονται από ευρωπαϊκές βάσεις δεδομένων, υπάρχει έλλειψη μετάφρασης των εντοπισμένων μεταβλητών κινδύνου σε άλλους μη ευρωπαϊκούς πληθυσμούς. Οι εναλλακτικές στρατηγικές που προτείνονται περιλαμβάνουν ανάλυση συνδέσεων. Πολύ πρόσφατα, η ταχέως μειούμενη τιμή πλήρους αλληλούχισης του γονιδιώματος έχει παράσχει επίσης μια ρεαλιστική εναλλακτική λύση για τις μελέτες GWAS που βασίζονται σε γονότυπα. Μπορεί να συζητηθεί εάν η χρήση αυτής της νέας τεχνικής εξακολουθεί να αναφέρεται ως μελέτη GWAS, αλλά η αλληλούχιση υψηλής απόδοσης έχει τη δυνατότητα να παράγει μερικά από τα μειονεκτήματα της μη ακολουθίας GWAS.[8]

2.3 Μοντέλα Κληρονομικότητας

Είναι σημαντικό να αναφερθούμε στους βασικούς νόμους της κληρονομικότητας για να κατανοήσουμε τον τρόπο μεταβίβασης των γονιδίων σε μια οικογένεια. Ένα ακριβές οικογενειακό ιατρικό ιστορικό, αποτελεί ένα πολύτιμο εργαλείο για την απεικόνιση του τρόπου με τον οποίο μεταδίδονται τα γονίδια από γενιά σε γενιά. Ένα άτομο έχει δύο αντίγραφα σχεδόν κάθε γονιδίου, ένα αντίγραφο από τη μητέρα του και ένα αντίγραφο από τον πατέρα του. Οι επιστήμονες έχουν μελετήσει τα ανθρώπινα γονίδια για να μάθουν πώς λειτουργούν κανονικά και πώς οι αλλαγές στα γονίδια μπορούν να αλλάξουν τον τρόπο λειτουργίας τους. Ορισμένες αλλαγές είναι πολύ μικρές και δεν επηρεάζουν τον τρόπο με τον οποίο λειτουργεί ένα γονίδιο. Αυτές οι αλλαγές συχνά ονομάζονται πολυμορφισμοί ενός νουκλεοτιδίου (SNPs, ή αλλιώς "snips") ή γονιδιακές παραλλαγές. Άλλες αλλαγές, που ονομάζονται μεταλλάξεις, επηρεάζουν το πώς λειτουργεί ένα γονίδιο και μπορεί να οδηγήσει σε ασθένειες. [9]

Σε κάποιες περιπτώσεις, τα μέλη μιας οικογένειας με την ίδια μετάλλαξη μπορεί να μην έχουν τα ίδια συμπτώματα. Για άλλες συνθήκες, τα άτομα με διαφορετικές μεταλλάξεις μπορούν να έχουν παρόμοια χαρακτηριστικά. Αυτό συμβαίνει επειδή η γονιδιακή έκφραση επηρεάζεται από γονίδια, καθώς και από το περιβάλλον. Οι ασθένειες που προκαλούνται από μεταλλάξεις σε ένα μόνο γονίδιο συνήθως κληρονομούνται με ένα απλό πρότυπο, ανάλογα με τη θέση του γονιδίου και αν χρειάζονται ένα ή δύο κανονικά αντίγραφα του γονιδίου. Αυτό συχνά αναφέρεται ως Μενδελική κληρονομικότητα, επειδή ο Gregor Mendel παρατήρησε για πρώτη φορά αυτά τα πρότυπα στα φυτά μπιζελιού.

Υπάρχουν κάποια βασικά μοντέλα κληρονομικότητας όπως είναι το Επικρατές(Dominant), το Υπολειπόμενο(Recessive) και το Συνεπικρατές(Codominant).

Για να εκφραστούν οι μεταλλάξεις του Επικρατούς μοντέλου αρκεί να υπάρχει μόνο ένα αντίγραφο αυτής της μετάλλαξης. Επομένως, όποιος κληρονομεί μία Επικρατή μετάλλαξη της νόσου, όπως η μετάλλαξη για τη νόσο του Huntington, θα έχει αυτή την ασθένεια. Οι επικρατείς κληρονομικές γενετικές ασθένειες τείνουν να εμφανίζονται σε κάθε γενιά μιας οικογένειας. Κάθε προσβεβλημένο άτομο έχει συνήθως έναν γονέα που έχει προσβληθεί. Ωστόσο, οι επικρατείς μεταλλάξεις μπορούν επίσης να συμβούν σε ένα άτομο για πρώτη φορά, χωρίς οικογενειακό ιστορικό της πάθησης (αυθόρμητη μετάλλαξη).

Οι υπολειπόμενες μεταλλάξεις απαιτούν δύο μεταλλαγμένα αντίγραφα για την εμφάνιση ασθένειας. Οι υπολειπόμενες γενετικές ασθένειες δεν εμφανίζονται συνήθως σε κάθε γενιά μιας οικογένειας που έχει προσβληθεί. Οι γονείς ενός προσβεβλημένου ατόμου είναι γενικά

φορείς: άτομα που δεν έχουν προσβληθεί και έχουν αντίγραφο μεταλλαγμένου γονιδίου. Αν και οι δύο γονείς είναι φορείς του ίδιου μεταλλαγμένου γονιδίου το μεταφέρουν και οι δύο στο παιδί, το παιδί θα επηρεαστεί.

Στο Συνεπικρατές μοντέλο τώρα, κανένα αλληλόμορφο δεν είναι υποχωρητικό και έτσι τελικώς θα εκφράστούν οι φαινότυποι και των δύο αλληλόμορφων.

Για να συμβολίσουμε τα αλληλόμορφα χρησιμοποιούμε είτε μικρά (υπολειπόμενο) και κεφαλαία(επικρατές) γράμματα π.χ. Αα το οποίο συμβολίζει ένα ετερόζυγο γονίδιο είτε δύο διαφορετικά γράμματα π.χ ΑΒ όπου το Α θα συμβολίζει το επικρατές αλληλόμορφο και το Β το υπολειπόμενο. Ο δεύτερος τρόπος είναι αυτός που προτιμούμε και εμείς στο εργαλείο μας για να συμβολίσουμε τους γονότυπους στα δεδομένα εισόδου.[10]

2.4 Μέτα-Ανάλυση & Συστηματική Ανασκόπηση

Η ανάγκη για έγκυρες και έγκαιρες αποφάσεις τόσο σε θέματα δημόσιας υγείας όσο και στην καθημερινή κλινική πρακτική, καθώς και η ολοένα αυξανόμενη πληροφορία σχετικά με τις διάφορες επιστημονικές υποθέσεις, καθιστούν απαραίτητη τη σύνθεση των αποτελεσμάτων που προέρχονται από την πληθώρα των μελετών που διεξάγονται. Για το σκοπό αυτόν, η πλέον αποδεκτή επιστημονική μέθοδος είναι η συστηματική ανασκόπηση(systematic review) σε συνδυασμό με την εφαρμογή της μετα-ανάλυσης (meta-analysis). Η συστηματική ανασκόπηση αποτελεί μια ανασκόπηση της βιβλιογραφίας σχετικά με μια συγκεκριμένη επιστημονική υπόθεση (π.χ. σχέση μεταξύ καπνισματικής συνήθειας και καρκίνου του πνεύμονα) και αποβλέπει στην αναγνώριση, την εκτίμηση και την επιλογή των καλύτερα μεθοδολογικά σχεδιασμένων μελετών.

Η μετα-ανάλυση αποτελεί μια μαθηματική διαδικασία που συνδυάζει στατιστικά τα αποτελέσματα των μελετών που επιλέχθηκαν έπειτα από τη συστηματική ανασκόπηση. Είναι προφανές ότι η εξαγωγή ασφαλών συμπερασμάτων με την εφαρμογή της μετα-ανάλυσης προϋποθέτει μια καλά σχεδιασμένη συστηματική ανασκόπηση της βιβλιογραφίας, έτσι ώστε να συμπεριληφθούν στη μετα-ανάλυση οι πλέον κατάλληλες μελέτες. Έτσι, τα συμπεράσματα της μετα-ανάλυσης είναι ασφαλή μόνον εφόσον έχει προηγηθεί ενδελεχής συστηματική ανασκόπηση όλων των μελετών, οι οποίες ενδεχομένως θα μπορούσαν να συμπεριληφθούν στη μετα-ανάλυση. Επιπλέον, η διεξαγωγή μόνο της συστηματικής ανασκόπησης χωρίς την εφαρμογή της μετα-ανάλυσης για τον υπολογισμό ενός συνδυαστικού αποτελέσματος με βάση τα ξεχωριστά αποτελέσματα των επιμέρους μελετών αποτελεί μια μη ολοκληρωμένη διαδικασία, καθώς δεν εξάγεται ένα συγκεντρωτικό αποτέλεσμα.

Γενικότερα, προτείνεται η σαφής διάκριση ανάμεσα στη συστηματική ανασκόπηση και τη μετα-ανάλυση, με την πρώτη να αποτελεί τη θεωρητική διαδικασία καθορισμού με ορισμένα κριτήρια των καλύτερα μεθοδολογικά σχεδιασμένων μελετών σχετικά με μια συγκεκριμένη επιστημονική υπόθεση και τη δεύτερη να αποτελεί τη μαθηματική διαδικασία υπολογισμού ενός συγκεντρωτικού αποτελέσματος με βάση τα αποτελέσματα των μελετών, οι οποίες επιλέχθηκαν έπειτα από τη συστηματική ανασκόπηση της βιβλιογραφίας. Έτσι, η συστηματική ανασκόπηση και η μετα-ανάλυση αποτελούν το πρώτο και το δεύτερο βήμα, αντίστοιχα, μιας διαδικασίας συνδυασμού των αποτελεσμάτων ενός αριθμού μελετών αναφορικά με μια συγκεκριμένη επιστημονική υπόθεση, με σκοπό τον υπολογισμό με μεγαλύτερη ακρίβεια και

εγκυρότητα σε σχέση με κάθε επιμέρους μελέτη ξεχωριστά ενός συγκεντρωτικού αποτελέσματος, το οποίο εκτιμά ουσιαστικά τη σχέση μεταξύ προσδιοριστή και έκβασης. Ουσιαστικά, η συστηματική ανασκόπηση και η μετα-ανάλυση αποτελούν δύο αλληλένδετες διαδικασίες και μόνον ο συνδυασμός τους μπορεί να οδηγήσει σε ασφαλή συμπεράσματα.

Σε ορισμένες περιπτώσεις, η μετα-ανάλυση αναφέρεται καταχρηστικά ως η διαδικασία που περιλαμβάνει τόσο τα βήματα μιας συστηματικής ανασκόπησης όσο και το στατιστικό συνδυασμό των αποτελεσμάτων των επιμέρους μελετών, ενώ σε ορισμένες άλλες περιπτώσεις αναφέρεται μόνο η εφαρμογή της συστηματικής ανασκόπησης, υπο-νοώντας όμως και την εφαρμογή της μετα-ανάλυσης. Είναι προτιμότερο πάντως να γίνεται διάκριση ανάμεσα στη συστηματική ανασκόπηση, ως της διαδικασίας αναζήτησης και επιλογής με ορισμένα κριτήρια των κατάλληλων μελετών, και στη μετα-ανάλυση, ως της στατιστικής διαδικασίας υπολογισμού ενός συγκεντρωτικού αποτελέσματος με βάση τα αποτελέσματα των επιμέρους μελετών. Στην πράξη, η μετα-ανάλυση είναι μια μαθηματική διαδικασία που χρησιμοποιεί τα αποτελέσματα της συστηματικής ανασκόπησης, ενώ η εξαγωγή ασφαλών συμπερασμάτων μέσω της μετα-ανάλυσης είναι αδύνατη χωρίς προηγουμένως να έχει πραγματοποιηθεί η κατάλληλη συστηματική ανασκόπηση.

Η συστηματική ανασκόπηση και η μετα-ανάλυση είναι απαραίτητες πριν από τη διεξαγωγή μιας νέας μελέτης, έτσι ώστε να είναι δυνατόν να διαπιστωθεί εάν τα αποτελέσματα της μελέτης αυτής προσθέτουν πληροφορία σχετικά με μια συγκεκριμένη επιστημονική υπόθεση. Η μετα-ανάλυση είναι περισσότερο χρήσιμη όταν τα αποτελέσματα των διαφόρων μελετών είναι αντιφατικά μεταξύ τους και όταν ο αριθμός των συμμετεχόντων στις μελέτες είναι μικρός, καθώς στις περιπτώσεις αυτές ο συνδυασμός των μελετών αυξάνει τη στατιστική ισχύ. Η μετα-ανάλυση αποτελεί ουσιαστικά μια στατιστική ανάλυση στην οποία ως «δεδομένα» χρησιμοποιούνται τα αποτελέσματα των επιμέρους μελετών και όχι τα πρωτογενή δεδομένα των μελετών αυτών. Στην περίπτωση πάντως που είναι διαθέσιμα και τα πρωτογενή δεδομένα των μελετών, συνιστάται η εφαρμογή της μεθόδου της συνένωσης (pooling) για τον υπολογισμό ενός συγκεντρωτικού αποτελέσματος. Η συνένωση είναι δυνατή σε ελάχιστες μόνο περιπτώσεις εξαιτίας της αδυναμίας συγκέντρωσης των πρωτογενών δεδομένων, αλλά παρουσιάζει αρκετά πλεονεκτήματα, με σημαντικότερο τη δυνατότητα εξουδετέρωσης των συγχυτών.[11]

2.4.1 Η μέτα-ανάλυση στην υγεία

Η μετα-ανάλυση, ως η μαθηματική διαδικασία συνδυασμού των αποτελεσμάτων διαφόρων μελετών για τον υπολογισμό ενός συγκεντρωτικού αποτελέσματος, εφαρμόστηκε για πρώτη φορά το 1904 από τον Καρλ Περσον σε μια προσπάθειά του να αντιμετωπίσει το πρόβλημα της μειωμένης στατιστικής ισχύος που παρουσίαζαν μελέτες με μικρό αριθμό συμμετεχόντων. Στις επιστήμες υγείας, η πρώτη μετα-ανάλυση δημοσιεύτηκε το 1955. Ο όρος μετα-ανάλυση (μετα-αναλυσίς) εισήχθη το 1976 από τον Γενε Γλας, το ενδιαφέρον του οποίου εστιαζόταν στην έρευνα στο χώρο της εκπαίδευσης.

Η μετα-ανάλυση στις επιστήμες υγείας, αρχικά, εφαρμόστηκε για την εκτίμηση της αποτελεσματικότητας θεραπευτικών παρεμβάσεων σε κλινικές δοκιμές. Η μετα-ανάλυση, για παράδειγμα, 33 κλινικών δοκιμών που εκτιμούσαν την αποτελεσματικότητα της ενδοφλέβιας χορήγησης στρεπτοκινάσης σε σχέση με τη χορήγηση ανενεργούς ουσίας (placebo) σε πάσχο-

ντες με έμφραγμα του μυοκαρδίου κατέδειξε την υπεροχή της στρεπτοκινάσης, ενώ μόνο οι 6 από τις 33 κλινικές δοκιμές είχαν καταλήξει σε στατιστικά σημαντική υπεροχή της στρεπτοκινάσης. Σήμερα, η δημοσίευση κλινικών δοκιμών απαιτεί την εφαρμογή αυστηρών κριτηρίων, διευκολύνοντας έτσι σημαντικά την εφαρμογή της μετα-ανάλυσης. Εξάλλου, στην περίπτωση της αιτιογνωστικής επιδημιολογίας, όπου δεν είναι δυνατή η πραγματοποίηση τυχαιοποιημένων μελετών, παρά μόνο η πραγματοποίηση μελετών «ασθενών-μαρτύρων» και σπανιότερα μελετών παρακολούθησης εφαρμόζονται όλο και πιο συχνά μετα-αναλύσεις για την ασφαλέστερη εκτίμηση της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης μιας πάθησης. Σημειώνεται, πάντως, ότι το πλεονέκτημα της τυχαιοποίησης που διαθέτουν οι κλινικές δοκιμές έναντι των μελετών «ασθενών-μαρτύρων» και των μελετών παρακολούθησης αντανakλάται και στην εφαρμογή της μετα-ανάλυσης στα διάφορα αυτά είδη μελετών. Μολοντί η μετα-ανάλυση μελετών που αφορούν στην αιτιογνωστική επιδημιολογία δεν είναι πάντοτε δυνατή (και ιδιαίτερα στις περιπτώσεις όπου σκοπός είναι ο υπολογισμός ενός συγκεντρωτικού μέτρου σχέσης), ο αριθμός των αντίστοιχων δημοσιεύσεων συνεχώς αυξάνεται, καθώς πριν από το 1992 είχαν δημοσιευτεί 678 μετα-αναλύσεις, ενώ μεταξύ των ετών 1992 και 1995 δημοσιεύτηκαν 525 και το 1996 δημοσιεύτηκαν 400.

2.4.2 Η μεθοδολογία της μέτα-ανάλυσης

Η μετα-ανάλυση εφαρμόζεται για το συνδυασμό, με μαθηματικό τρόπο, των αποτελεσμάτων των μελετών που επιλέχθηκαν κατά τη συστηματική βιβλιογραφική ανασκόπηση, έτσι ώστε να εξαχθεί ένα συγκεντρωτικό αποτέλεσμα. Για παράδειγμα, εάν σε μια συστηματική ανασκόπηση επιλέχθηκαν τρεις μελέτες με 40, 60 και 100 συμμετέχοντες σε καθεμία, τότε η μετα-ανάλυση των τριών αυτών μελετών θα εκφράζει ουσιαστικά το αποτέλεσμα μιας μελέτης με 200 συμμετέχοντες. Με τον τρόπο αυτόν, η μετα-ανάλυση οδηγεί στον υπολογισμό ενός συγκεντρωτικού αποτελέσματος με μεγαλύτερη ακρίβεια και εγκυρότητα απ' ό,τι κάθε μελέτη χωριστά. Η μετα-ανάλυση παρέχει την καλύτερη δυνατή ένδειξη (evidence) όταν περιλαμβάνει τυχαιοποιημένες ελεγχόμενες δοκιμές, ενώ παρουσιάζει μειωμένη εγκυρότητα στην περίπτωση μελετών παρακολούθησης και μελετών «ασθενών-μαρτύρων».[12]

Σφάλμα Δημοσίευσης

Πριν από την πραγματοποίηση της μετα-ανάλυσης απαιτείται η εκτίμηση του σφάλματος δημοσίευσης (publication bias), το οποίο μπορεί να οφείλεται στο γεγονός ότι:

- Τα περισσότερα περιοδικά δημοσιεύουν συχνότερα μελέτες οι οποίες καταλήγουν σε θετικά ευρήματα έναντι μελετών που καταλήγουν σε αρνητικά ευρήματα, δημοσιεύουν δηλαδή συχνότερα μελέτες που καταλήγουν στο ότι η ενδεικτική κατηγορία του μελετώμενου προσδιοριστή αυξάνει τη συχνότητα εμφάνισης της μελετώμενης έκβασης (π.χ. μελέτες που καταλήγουν στο συμπέρασμα ότι οι καπνιστές εμφανίζουν συχνότερα καρκίνο του πνεύμονα).
- Οι ερευνητές αποστέλλουν συχνότερα προς κρίση στα περιοδικά μελέτες που καταλήγουν σε θετικά ευρήματα έναντι μελετών που καταλήγουν σε αρνητικά ευρήματα

- Οι μελέτες που καταλήγουν σε θετικά ευρήματα δημο-σιεύονται συχνότερα σε περιοδικά τα οποία εκδίδονται στην αγγλική γλώσσα, με αποτέλεσμα να έχουν μεγαλύτερη πιθανότητα να εντοπιστούν από τους ερευνητές που διεξάγουν μια συστηματική ανασκόπηση.
- Οι μελέτες που καταλήγουν σε θετικά ευρήματα έχουν μεγαλύτερη πιθανότητα να εντοπιστούν κατά τη διαδικασία της συστηματικής ανασκόπησης.
- Οι μελέτες που καταλήγουν σε θετικά ευρήματα έχουν μεγαλύτερη πιθανότητα να δημοσιευτούν σε περισσότερα από ένα περιοδικά, με αποτέλεσμα να έχουν και μεγαλύτερη πιθανότητα να συμπεριληφθούν σε μια συστηματική ανασκόπηση.
- Ορισμένες μελέτες δεν αποστέλλονται ποτέ για δημοσίευση και ιδιαίτερα εκείνες που δεν καταλήγουν σε θετικά ευρήματα, εκείνες που χρηματοδοτούνται από φαρμακευτικές εταιρείες και δεν καταλήγουν στα επιθυμητά αποτελέσματα και εκείνες που αφορούν σε μεταπτυχιακές ή διδακτορικές εργασίες.

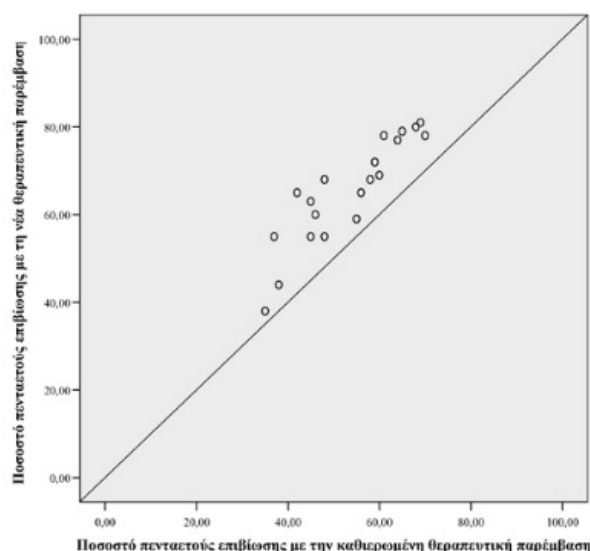
Η ύπαρξη σφάλματος δημοσίευσης αποτελεί συστηματικό σφάλμα που μειώνει σημαντικά την εγκυρότητα μιας μετα-ανάλυσης και για το λόγο αυτόν απαιτείται ο εντοπισμός του.^[13]

Ετερογένεια των μελετών

Η ακρίβεια και η εγκυρότητα μιας μετα-ανάλυσης εξαρτώνται σημαντικά από το βαθμό στον οποίο οι επιμέρους μελέτες είναι αρκετά ομοιογενείς μεταξύ τους, έτσι ώστε τα αποτελέσματά τους να μπορούν να συνδυαστούν για τον υπολογισμό ενός συγκεντρωτικού αποτελέσματος. Έτσι, πρέπει να υπάρχει ομοιογένεια στο μεθοδολογικό σχεδιασμό, στους μελετώμενους πληθυσμούς, στον τρόπο μέτρησης του προσδιοριστή και της έκβασης, στις μεθόδους εξουδετέρωσης των συγχυτών κ.λπ. Σε κάθε περίπτωση, βέβαια, τα αποτελέσματα των επιμέρους μελετών είναι λογικό να παρουσιάζουν μια ορισμένη μεταβλητότητα που οφείλεται στην τύχη. Όταν όμως τα αποτελέσματα των επιμέρους μελετών που πρόκειται να συμπεριληφθούν στη μετα-ανάλυση παρουσιάζουν μεγαλύτερη ετερογένεια από εκείνη που αναμένεται εκ τύχης, τότε ο υπολογισμός ενός μόνο συγκεντρωτικού αποτελέσματος μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα. Στην περίπτωση αυτή, συνιστάται η εφαρμογή της διαστρωματικής ανάλυσης ή της ανάλυσης παλινδρόμησης με τη χρήση των διαφόρων χαρακτηριστικών του μεθοδολογικού σχεδιασμού των μελετών ως προβλεπτικών παραγόντων για την εκτίμηση του βαθμού στον οποίο τα χαρακτηριστικά αυτά επηρεάζουν τη σχέση μεταξύ προσδιοριστή και έκβασης. Οι μελέτες των οποίων τα αποτελέσματα διαφέρουν σημαντικά από τα αποτελέσματα της πλειονότητας των μελετών δεν πρέπει να απορρίπτονται απλά και μόνο εξαιτίας της ασυμφωνίας αυτής, αλλά πρέπει να εξετάζονται διεξοδικά τα διάφορα χαρακτηριστικά του μεθοδολογικού σχεδιασμού ή της ανάλυσης των δεδομένων που μπορεί να οδήγησαν στην ασυμφωνία αυτή.

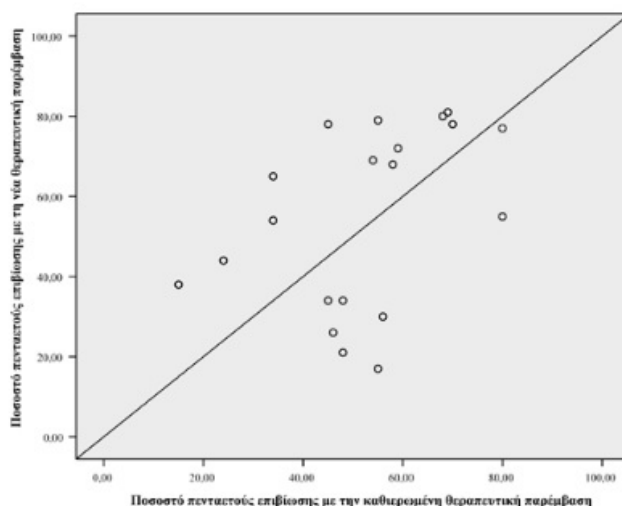
Η εκτίμηση της ομοιογένειας των αποτελεσμάτων των μελετών που πρόκειται να συμπεριληφθούν στη μετα-ανάλυση μπορεί να πραγματοποιηθεί είτε με το διάγραμμα L' Abbé (L'Abbé plot) είτε με την εφαρμογή των κατάλληλων στατιστικών ελέγχων. Στο διάγραμμα L' Abbé, στον κάθετο άξονα αντιστοιχεί το μέτρο συχνότητας για εκείνους που ανήκουν στην

ενδεικτική κατηγορία του μελετώμενου προσδιοριστή, ενώ στον οριζόντιο άξονα αντιστοιχεί το μέτρο συχνότητας για εκείνους που ανήκουν στην κατηγορία αναφοράς του προσδιοριστή. Διεξάγεται μια μετα-ανάλυση, π.χ., για την εκτίμηση της αποτελεσματικότητας μιας νέας θεραπευτικής παρέμβασης για την αντιμετώπιση του καρκίνου του μαστού. Το μέτρο συχνότητας που υπολογίζεται είναι το ποσοστό πενταετούς επιβίωσης. Στην παρακάτω εικόνα φαίνεται το διάγραμμα L' Abbé της μετα-ανάλυσης αυτής. Η διαγώνιος γραμμή που φέρεται από την αρχή των αξόνων σε γωνία 45 αποτελεί το όριο της αποτελεσματικότητας της νέας θεραπευτικής παρέμβασης, με τις τιμές που βρίσκονται πάνω από τη διαγώνιο αυτή να εκφράζουν την υπεροχή της νέας θεραπευτικής παρέμβασης έναντι της καθιερωμένης και τις τιμές που βρίσκονται κάτω από τη διαγώνιο να εκφράζουν την υπεροχή της καθιερωμένης θεραπευτικής παρέμβασης. Όσο πιο συμπαγές είναι το διάγραμμα L' Abbé, όσο πιο μεγάλη συγκέντρωση δηλαδή παρουσιάζουν οι διάφορες τιμές, τόσο πιο ομοιογενείς είναι οι μελέτες από τις οποίες προέκυψαν τα δεδομένα του διαγράμματος. Στην εικόνα 2.3 οι τιμές του διαγράμματος παρουσιάζουν μεγάλη συγκέντρωση, η οποία υποδηλώνει ομοιογένεια των μελετών, ενώ το αντίστροφο συμβαίνει στην αμέσως επόμενη εικόνα(2.4).[14]



Εικόνα 2.3: Διάγραμμα L' Abbé, με τις τιμές που αντιστοιχούν στις 20 μελέτες να παρουσιάζουν μεγάλη συγκέντρωση, γεγονός που υποδηλώνει την ύπαρξη ομοιογένειας.

Η ερμηνεία του διαγράμματος L' Abbé εναπόκειται στην υποκειμενική κρίση των ερευνητών και απαιτεί ιδιαίτερη εμπειρία για την αποφυγή λανθασμένων συμπερασμάτων. Για το λόγο αυτόν, το διάγραμμα L'Abbé χρησιμοποιείται επικουρικά για την εκτίμηση της ομοιογένειας των μελετών και τα συμπεράσματα στηρίζονται κυρίως στην εφαρμογή των κατάλληλων στατιστικών ελέγχων. Στην περίπτωση αυτή, συνήθως προτιμάται η χρήση του στατιστικού ελέγχου Q των DerSimonian και Laird, που παρουσιάζει τη μεγαλύτερη εγκυρότητα και στατιστική ισχύ και επιπλέον υπολογίζεται απλούστερα. Η μηδενική υπόθεση που ελέγχεται είναι ότι οι μελέτες είναι ομοιογενείς μεταξύ τους, οπότε εάν το παρατηρούμενο επίπεδο της στατιστικής σημαντικότητας (τιμή p) είναι μεγαλύτερο από το προκαθορισμένο επίπεδο της στατιστικής σημαντικότητας (τιμή α), τότε δεν απορρίπτεται η μηδενική υπόθεση και, επο-



Εικόνα 2.4: Διάγραμμα L' Abbé, με τις τιμές που αντιστοιχούν στις 20 με-βιέτες να παρουσιάζουν μικρή συγκέντρωση, γεγονός που υποδηλώνει την ύπαρξη ετερογένειας.

μένως, υπάρχει ομοιογένεια μεταξύ των μελετών. Αντίθετα, εάν η τιμή p που προκύπτει από την εφαρμογή του στατιστικού ελέγχου είναι μικρότερη από την τιμή α , τότε απορρίπτεται η μηδενική υπόθεση και υπάρχει ετερογένεια μεταξύ των μελετών. Η εκτίμηση της ύπαρξης ετερογένειας είναι εξαιρετικής σημασίας, καθώς καθορίζει τη μαθηματική μέθοδο που πρόκειται να εφαρμοστεί για τον υπολογισμό του συγκεντρωτικού αποτελέσματος. [14]

Η Μαθηματική Διαδικασία

Η εκτίμηση του σφάλματος δημοσίευσης και της ομοιογένειας μεταξύ των μελετών αποτελούν απαραίτητες προϋποθέσεις για την εφαρμογή της κατάλληλης μαθηματικής μεθόδου, με σκοπό τον υπολογισμό του συγκεντρωτικού αποτελέσματος της μετα-ανάλυσης. Εάν διαπιστωθεί ομοιογένεια μεταξύ των μελετών, τότε εφαρμόζεται το μοντέλο των σταθερών επιδράσεων (fixed-effects model), ενώ εάν διαπιστωθεί ετερογένεια μεταξύ των μελετών, τότε εφαρμόζεται το μοντέλο των τυχαίων επιδράσεων (random-effects model). Στην περίπτωση της ύπαρξης ομοιογένειας, η επιλογή του μοντέλου δεν επηρεάζει τα αποτελέσματα.

Όταν εφαρμόζεται το μοντέλο των σταθερών επιδράσεων, ισχύει η υπόθεση ότι το αποτέλεσμα του προσδιοριστή είναι σταθερό στις διάφορες μελέτες και ότι η μεταβλητότητα του αποτελέσματος οφείλεται στην τυχαία μεταβλητότητα που παρουσιάζει η κάθε μελέτη εξαιτίας του γεγονότος ότι λαμβάνεται ένα «δείγμα» από τον πληθυσμό πηγή. Επιπλέον, όταν εφαρμόζεται το μοντέλο των σταθερών επιδράσεων, τότε τα συμπεράσματα της μετα-ανάλυσης είναι έγκυρα μόνο για τις μελέτες που συμπεριελήφθησαν στην ανάλυση αυτή και δεν μπορούν να γενικευτούν για όλες τις παρόμοιες μελέτες. Και στα δύο μοντέλα, έχουμε ως στόχο να αναθέσουμε περισσότερο βάρος στις έρευνες που περιέχουν την περισσότερη πληροφορία. Η ανάθεση αυτή γίνεται στο μοντέλο των σταθερών επιδράσεων χρησιμοποιώντας την αντίστροφη διακύμανση καθώς είναι πιο ακριβής σε ότι έχει να κάνει με το πραγματικό βάρος της κάθε έρευνας και βοηθάει στο να μειωθεί η διακύμανση της συνολικής επίδρασης. Έτσι

ο τύπος που χρησιμοποιείται είναι ο εξής :

$$w_i = \frac{1}{v_i}$$

και όπου v_i είναι η διακύμανση της έρευνας i . Ο σταθμισμένος μέσος της μετα-ανάλυσης υπολογίζεται κατόπιν ως :

$$T. = \frac{\sum_{i=1}^k w_i T_i}{\sum_{i=1}^k w_i}$$

το οποίο εκφράζει το πηλίκο του αθροίσματος όλων των Odds Ratio σταθμισμένα με το βάρος της εκάστοτε έρευνας προς το άθροισμα όλων των βαρών. Η διακύμανση της συνδυασμένης επίδρασης περιγράφεται απ' τον τύπο :

$$v. = \frac{1}{\sum_{i=1}^k w_i}$$

. Στη συνέχεια υπολογίζουμε το τυπικό σφάλμα της συνδυασμένης επίδρασης (Standard Error):

$$SE = \sqrt{v.}$$

που είναι η ρίζα της συνδυασμένης διακύμανσης. Το διάστημα εμπιστοσύνης 95 % της συνδυασμένης επίδρασης υπολογίζεται ως

$$LowerLimit = T. - 1.96 * SE(T.)$$

$$UpperLimit = T. + 1.96 * SE(T.)$$

Τελικώς για να υπολογίσουμε την τιμή Z :

$$Z = \frac{T.}{SE(T.)}$$

η οποία στην συνέχεια είναι απαραίτητη για τον υπολογισμό του P-Value που υπολογίζεται από τον τύπο :

$$p = 2 - (1 - \phi(|Z|))$$

όπου $\Phi(Z)$ είναι μία τιμή της συνάρτησης κανονικής κατανομής.

Το μοντέλο σταθερού αποτελέσματος, που συζητήθηκε παραπάνω, ξεκινά με την παραδοχή ότι το πραγματικό αποτέλεσμα είναι το ίδιο σε όλες τις μελέτες. Ωστόσο, αυτή η υπόθεση μπορεί να μην είναι εφικτή σε πολλές συστηματικές ανασκοπήσεις. Όταν αποφασίζουμε να ενσωματώσουμε μια ομάδα μελετών σε μια μετα-ανάλυση, υποθέτουμε ότι οι μελέτες έχουν αρκετά κοινά, συνεπώς υπάρχει νόημα να συνθέτουν τις πληροφορίες. Ωστόσο, δεν υπάρχει λόγος να υποθέσουμε ότι είναι "ταυτόσημα" με την έννοια ότι το πραγματικό μέγεθος αποτελέσματος είναι ακριβώς το ίδιο σε όλες τις μελέτες.

Στο πλαίσιο του μοντέλου τυχαίων επιδράσεων πρέπει να λάβουμε υπόψη δύο επίπεδα δειγματοληψίας και δύο πηγές σφαλμάτων. Πρώτον, τα πραγματικά μεγέθη επίδρασης θ , διανέμονται γύρω από μ με διακύμανση τ^2 που αντανakλά την πραγματική κατανομή των πραγματικών αποτελεσμάτων ως προς το μέσο όρο τους. Δεύτερον, το παρατηρούμενο απο-

τέλεσμα T για κάθε δεδομένο θ θα κατανεμηθεί γύρω από το θ με μια διακύμανση σ^2 που εξαρτάται κυρίως από το μέγεθος του δείγματος για αυτή τη μελέτη. Επομένως, κατά την εκχώρηση βαρών στις έρευνες για την εκτίμηση μ , πρέπει να αντιμετωπίσουμε και τις δύο πηγές δειγματοληπτικού σφάλματος - μέσα στις μελέτες (ϵ) και μεταξύ των μελετών (ζ).

Η προσέγγιση μιας ανάλυσης τυχαίων επιδράσεων είναι να αποσυντεθεί η παρατηρούμενη διακύμανση στα δύο συνιστώμενα μέρη της, εντός των μελετών και μεταξύ των μελετών, και στη συνέχεια να χρησιμοποιηθούν και τα δύο μέρη κατά την εκχώρηση των βαρών. Ο στόχος θα είναι να ληφθούν υπόψη και οι δύο πηγές αβεβαιότητας. Ο μηχανισμός που χρησιμοποιείται για την αποσύνθεση της διακύμανσης είναι ο υπολογισμός της ολικής διακύμανσης (που παρατηρείται) και στη συνέχεια η απομόνωση της διακύμανσης εντός των μελετών. Η διαφορά μεταξύ αυτών των δύο τιμών θα μας δώσει τη διαφορά μεταξύ των μελετών, η οποία ονομάζεται tau-squared (τ^2).

Θα υπολογίσουμε το Q , το οποίο αντιπροσωπεύει τη συνολική διακύμανση και το df , το οποίο αντιπροσωπεύει την αναμενόμενη διακύμανση εάν όλες οι μελέτες έχουν την ίδια πραγματική επίδραση. Η διαφορά, $Q - df$, θα μας δώσει την περίσσεια διακύμανση. Τέλος, η τιμή αυτή θα μετασχηματιστεί, για να την τοποθετήσει στην ίδια κλίμακα με τη διακύμανση εντός της μελέτης. Αυτή η τελευταία τιμή ονομάζεται tau-squared (τ^2).

Η στατιστική Q αντιπροσωπεύει τη συνολική διακύμανση και ορίζεται ως

$$Q = \sum_{i=1}^k w_i (T_i - T.)^2$$

δηλαδή το άθροισμα των τετραγωνικών αποκλίσεων κάθε μελέτης (T_i) από το συνδυασμένο μέσο όρο ($T.$). Σημειώστε το " w_i " στον τύπο, που δείχνει ότι κάθε μία από τις τετραγωνικές αποκλίσεις σταθμίζεται από την αντίστροφη διακύμανση της μελέτης. Μια μεγάλη μελέτη που απέχει πολύ από τον μέσο όρο θα έχει μεγαλύτερο αντίκτυπο στο Q από μια μικρή μελέτη στην ίδια τοποθεσία. Μια ισοδύναμη φόρμουλα, χρήσιμη για υπολογισμούς, είναι

$$Q = \sum_{i=1}^k w_i T_i^2 - \frac{(\sum_{i=1}^k w_i T_i)^2}{\sum_{i=1}^k w_i}$$

Δεδομένου ότι το X αντανακλά τη συνολική διακύμανση, πρέπει τώρα να αναλυθεί στα συστατικά μέρη του. Εάν η μόνη πηγή διακύμανσης ήταν σφάλμα στο εσωτερικό της μελέτης, τότε η αναμενόμενη τιμή του X θα είναι οι βαθμοί ελευθερίας (df) για τη μετα-ανάλυση όπου

$$df = \text{Number of Studies} - 1$$

Αυτό μας επιτρέπει να υπολογίσουμε τη διαφορά μεταξύ των μελετών, τ^2 , ως εξής:

$$\tau^2 = \begin{cases} \frac{Q-df}{C} & \text{if } Q > df \\ 0 & \text{if } Q < df \end{cases}$$

όπου

$$C = \sum_{i=1} w_i - \frac{\sum_{i=1} w_i^2}{\sum_{i=1} w_i}$$

Ο αριθμητής, $Q - df$, είναι η περίσσεια (παρατηρούμενη μείον αναμενόμενη) διακύμανση. Ο παρονομαστής, C , είναι ένας παράγοντας κλιμάκωσης που έχει να κάνει με το γεγονός ότι το X είναι ένα σταθμισμένο άθροισμα τετραγώνων. Με την εφαρμογή αυτού του συντελεστή κλιμάκωσης εξασφαλίζουμε ότι το tau-squared είναι στην ίδια μετρική με τη διακύμανση εντός των μελετών.

Στην ανάλυση τυχαίων επιδράσεων, κάθε μελέτη θα σταθμιστεί από το αντίστροφο της διακύμανσής της. Η διαφορά είναι ότι η διακύμανση περιλαμβάνει πλέον την αρχική διακύμανση μεταξύ των μελετών και τη διακύμανση μεταξύ των μελετών, tau-squared. Χρησιμοποιούμε τις ίδιες συναρτήσεις, αλλά προσθέτουμε ένα (*) για να αναπαραστήσουμε το μοντέλο των τυχαίων επιδράσεων. Συγκεκριμένα, κάτω από το πρότυπο τυχαίων επιδράσεων το βάρος που αποδίδεται σε κάθε μελέτη είναι:

$$w_{i*} = \frac{1}{v_{i*}}$$

όπου v_{i*} είναι η διακύμανση εντός της μελέτης για τη μελέτη (i) συν τη διακύμανση μεταξύ των μελετών, tau-squared. Αυτό είναι,

$$v_{i*} = v_i + \tau^2$$

Ο σταθμισμένος μέσος στη συνέχεια υπολογίζεται ως:

$$T_{.*} = \frac{\sum_{i=1}^k w_i * T_i}{\sum_{i=1}^k w_{i*}}$$

δηλαδή το άθροισμα των προϊόντων (μέγεθος επίδρασης πολλαπλασιασμένο επί το βάρος) διαιρούμενο με το άθροισμα των βαρών. Η διακύμανση του συνδυασμένου αποτελέσματος ορίζεται ως η αμοιβαιότητα του αθροίσματος των βαρών ή

$$v_{.*} = \frac{1}{\sum_{i=1}^k w_{i*}}$$

και το συνολικό σφάλμα της ολικής επίδρασης είναι η ρίζα της διακύμανσης

$$SE = \sqrt{v_{.*}}$$

Το διάστημα εμπιστοσύνης 95 % της συνδυασμένης επίδρασης υπολογίζεται ως

$$LowerLimit* = T_{.*} - 1.96 * SE(T_{.*})$$

$$UpperLimit* = T_{.*} + 1.96 * SE(T_{.*})$$

Τελικώς για να υπολογίσουμε την τιμή Z :

$$Z^* = \frac{T.^*}{SE(T.^*)}$$

η οποία στην συνέχεια είναι απαραίτητη για τον υπολογισμό του P-Value που υπολογίζεται από τον τύπο :

$$p^* = 2 - (1 - \phi(|Z^*|))$$

όπου $\Phi(Z^*)$ είναι μία τιμή της συνάρτησης κανονικής κατανομής.

Στις περισσότερες περιπτώσεις μια μέτα-ανάλυση συνοδεύεται από ένα διάγραμμα το οποίο ονομάζεται forest plot πάνω στο οποίο αναπαριστώνται τα αποτελέσματα τόσο των επιμέρους μελετών όσο και της μέτα-ανάλυσης. Η σημαντικότητα του διαγράμματος αυτού έγκειται στο γεγονός ότι δύναται να προσφέρει μια άμεση και σαφή εικόνα των αποτελεσμάτων της μέτα-ανάλυσης, ενώ η ερμηνεία του δεν απαιτεί εξειδικευμένες γνώσεις.[15]

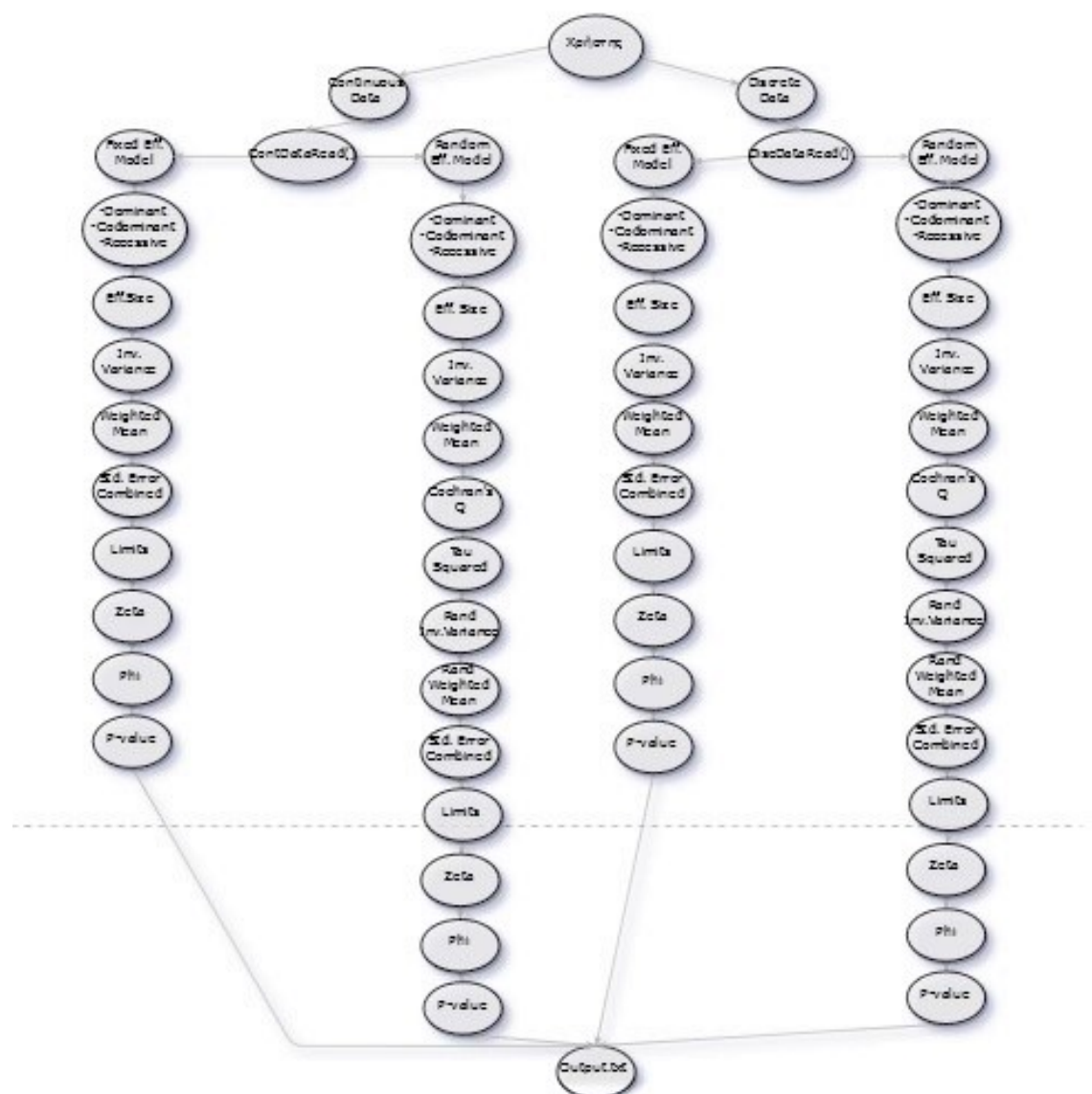
Μέρος **III**

Πρακτικό Μέρος

Κεφάλαιο 3

Περιγραφή Λογισμικού

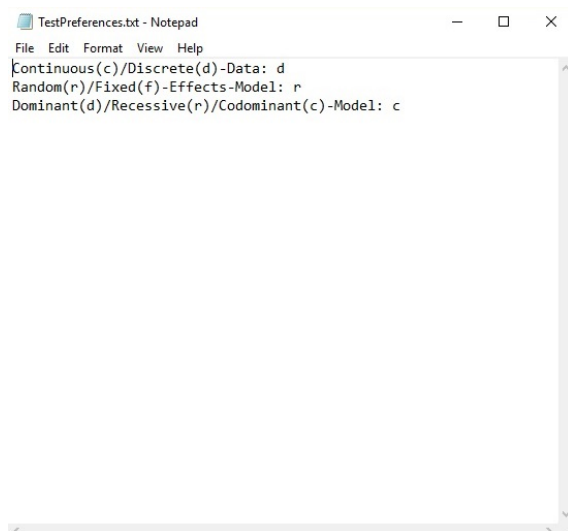
Στο κεφάλαιο αυτό γίνεται μια περιγραφή της λειτουργίας του λογισμικού σε επίπεδο ανάπτυξης κώδικα .



Εικόνα 3.1: Διάγραμμα Ροής του Λογισμικού GWASMetAnalysis

3.1 Δεδομένα εισόδου

Καταρχάς, πρώτου ξεκινήσει η διαδικασία των υπολογισμών ο χρήστης οφείλει να παρέχει στο σύστημα τα δεδομένα με μία συγκεκριμένη δομή.



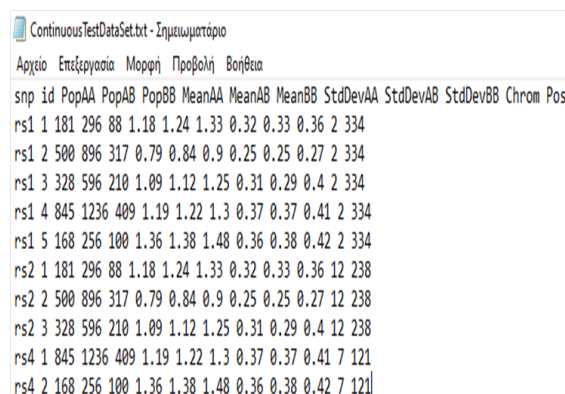
Εικόνα 3.2: Δομή αρχείου επιλογών χρήστη

Αρχείο	Επεξεργασία	Μορφή	Προβολή	Βοήθεια					
snp id	aa1	ab1	bb1	aa0	ab0	bb0	Chrom	Pos	
rs1	1	40	587	1325	72	684	2180	1	345
rs1	2	97	620	1100	88	579	1958	1	345
rs1	3	37	343	1256	64	589	2001	1	345
rs1	4	63	445	1182	94	644	1991	1	345
rs1	5	72	680	1005	78	499	2067	1	345
rs2	1	40	587	1325	72	684	2180	5	987
rs2	2	97	620	1100	88	579	1958	5	987
rs2	3	37	343	1256	64	589	2001	5	987
rs3	1	63	445	1182	94	644	1991	7	243
rs3	2	72	680	1005	78	499	2067	7	243
rs1	1	40	587	1325	72	684	2180	1	345
rs1	2	97	620	1100	88	579	1958	1	345
rs1	3	37	343	1256	64	589	2001	1	345
rs1	4	63	445	1182	94	644	1991	1	345
rs1	5	72	680	1005	78	499	2067	1	345
rs2	1	40	587	1325	72	684	2180	5	987
rs2	2	97	620	1100	88	579	1958	5	987
rs2	3	37	343	1256	64	589	2001	5	987
rs3	1	63	445	1182	94	644	1991	7	243
rs3	2	72	680	1005	78	499	2067	7	243
rs1	1	40	587	1325	72	684	2180	1	345
rs1	2	97	620	1100	88	579	1958	1	345
rs1	3	37	343	1256	64	589	2001	1	345

Εικόνα 3.3: Δομή αρχείου διακριτών δεδομένων εισόδου χρήστη

Αρχικά μέσω του αρχείου Preferences ο χρήστης καλείται να επιλέξει μέσω ποιων μοντέλων επιθυμεί να εξαχθούν τα αποτελέσματα του. Όπως γίνεται κατανοητό και στην εικόνα 3.2, πρώτα επιλέγει ποιο τύπο δεδομένων θα παρέχει στο λογισμικό (Συνεχή-c ή Διακριτά-d), στη συνέχεια αν θα χρησιμοποιήσει το μοντέλο των σταθερών ή των τυχαίων επιδράσεων(Σταθερά-f ή Τυχαία-r), ενώ τέλος επιλέγει μεταξύ ποιου μοντέλου κληρονομικότητας αλληλομόρφου(Επικρατές-d, Συνεπικρατές-c, Υπολοιπόμενο-r) επιθυμεί να χρησιμοποιήσει.

Κατόπιν, ανάλογα με την πρώτη του επιλογή δηλαδή ποιον τύπο δεδομένων θα χρησιμοποιήσει, πρέπει να παρέχει τα ανάλογα δεδομένα. Εάν επιλέξει τα διακριτά δεδομένα, η δομή



snp	id	PopAA	PopAB	PopBB	MeanAA	MeanAB	MeanBB	StdDevAA	StdDevAB	StdDevBB	Chrom	Pos
rs1	1	181	296	88	1.18	1.24	1.33	0.32	0.33	0.36	2	334
rs1	2	500	896	317	0.79	0.84	0.9	0.25	0.25	0.27	2	334
rs1	3	328	596	210	1.09	1.12	1.25	0.31	0.29	0.4	2	334
rs1	4	845	1236	409	1.19	1.22	1.3	0.37	0.37	0.41	2	334
rs1	5	168	256	100	1.36	1.38	1.48	0.36	0.38	0.42	2	334
rs2	1	181	296	88	1.18	1.24	1.33	0.32	0.33	0.36	12	238
rs2	2	500	896	317	0.79	0.84	0.9	0.25	0.25	0.27	12	238
rs2	3	328	596	210	1.09	1.12	1.25	0.31	0.29	0.4	12	238
rs4	1	845	1236	409	1.19	1.22	1.3	0.37	0.37	0.41	7	121
rs4	2	168	256	100	1.36	1.38	1.48	0.36	0.38	0.42	7	121

Εικόνα 3.4: Δομή αρχείου συνεχών δεδομένων εισόδου χρήστη

του αρχείου του πρέπει να συμμορφώνεται σύμφωνα με αυτή της εικόνας 3.3. Στην πρώτη στήλη καταγράφονται τα ονόματα των ερευνών ενώ στις υπόλοιπες οι πληθυσμοί control(0) και case(1). Σημειώνεται ότι όλα πρέπει να έχουν 1 κενό μεταξύ τους. Στην περίπτωση των συνεχών δεδομένων, ο χρήστης ακολουθεί την εικόνα 3.4. Στην πρώτη στήλη πάλι δίδονται δεδομένα για την ονομασία των επιμέρους ερευνών ενώ στις υπόλοιπες πρέπει να δωθούν δεδομένα για τους γονοτύπους του συνολικού πληθυσμού (Πλήθος Ατόμων με AA, AB, BB), το μέσο όρο της κάθε κατηγορίας καθώς και την τυπική της απόκλιση.

3.2 Ανάλυση Κλάσεων

Στην παρούσα υποενότητα θα αναλυθεί ενδελεχώς η υλοποίηση των επιμέρους κλάσεων που συνθέτουν το υπο συζήτηση λογισμικό.

3.2.1 DataRead.java

Εντός της παρούσας κλάσης γίνεται ορισμός λιστών για την αποθήκευση των δεδομένων που εισάγονται από τους χρήστες. Η κλάση αυτή περιέχει δύο συναρτήσεις, την DiscreteDataRead() και την ContinuousDataRead(), οι οποίες χρησιμοποιούνται για διάβασμα σταθερών ή συνεχών δεδομένων αντίστοιχα και για την αποθήκευση των δεδομένων αυτών για μελλοντική χρήση από τις υπόλοιπες κλάσεις του λογισμικού. Η επιλογή μεταξύ των δύο συναρτήσεων, γίνεται αυτόματα από το λογισμικό ανάλογα με την επιλογή που έχει διαβαστεί από το αρχείο επιλογών του χρήστη.

3.2.2 MetaStats.java

Στην MetaStats.java υλοποιούνται όλες οι χρήσιμες στατιστικές συναρτήσεις για την διεξαγωγή της μέτα-ανάλυσης κυρίως για τα διακριτά δεδομένα. Πολλές συναρτήσεις όμως είναι κοινές και στους δύο τύπους δεδομένων όπως γίνεται αντιληπτό και στο άνωθεν flowchart. Οι Codom-, Dom- και Rec-OddsRatio() χρησιμοποιούνται για την εύρεση του odds ratio με βάση το συνεπικρατές, το επικρατές και το υπολειπόμενο μοντέλο αντίστοιχα. Οι Codom-, Dom- και Rec-FixedInvVariance() χρησιμοποιούνται για τον υπολογισμό της αντίστροφης διακύμανσης στο μοντέλο σταθερών επιδράσεων (Fixed Effects Μοντέλο) για κάθε επιμέρους

μοντέλο(όπως αναλύθηκαν παραπάνω) ξεχωριστά. Η `WeightedMean()` υπολογίζει το σταθμισμένο μέσο των ερευνών, ενώ η `StdErrorCombined()` το συνολικό τυπικό σφάλμα `Standard Error` που προκύπτει από τις έρευνες. Έπειτα μέσω της `Limits` εντοπίζονται τα όρια για διάστημα εμπιστοσύνης 95% και μέσω της `Zeta` υπολογίζουμε τον αριθμό που εισέρχεται στη συνάρτηση $\Phi(Z)$. Τέλος, μέσω της `Pvalue()` υπολογίζεται το τελικό p-value που θα δοθεί στο χρήστη. Οι συναρτήσεις `CochransQ()`, `TauSquared()` και `RandomInvVariance()` χρησιμοποιούνται συνδυαστικά με τις προηγούμενες συναρτήσεις, όταν ο χρήστης επιλέγει να εκτελέσει τους υπολογισμούς του με το `Random Effects Model`. Οι τρεις αυτές συναρτήσεις υπολογίζουν το Q του Cochran, το τ^2 και τέλος χρησιμοποιώντας αυτά με την `RandomInvVariance()` προσθέτουν την επιπλέον διακύμανση στη `FixedInvVariance()`.

3.2.3 ContMetaStats.java

Η μοναδική διαφορά που υπάρχει εν συγκρίσει με την απλή `MetaStats` είναι πως εδώ χειριζόμαστε συνεχή δεδομένα. Μέσω των `Codom-`, `Dom-` και `Rec-EffSize()` υπολογίζουμε το effect size για κάθε ένα εκ των μοντέλων (χρησιμοποιείται αντι για την odds ratio) και εν συνεχεία υπολογίζουμε το την αντίστροφη διακύμανση με τις υπόλοιπες τρεις συναρτήσεις. Τα αποτελέσματα που προκύπτουν χρησιμοποιούνται από τις συναρτήσεις της `MetaStats`, δηλαδή για τις υπόλοιπες μετρικές δεν υπάρχει καμία περαιτέρω αλλαγή.

3.2.4 PreferenceHandling.java

Αρχικά μέσω της `FileReading()` διαβάζουμε το αρχείο με τις οδηγίες που μας δίνει ο χρήστης. Αφού αποθηκεύουμε τις επιλογές περνάμε στην `Choice()` όπου ανοίγουμε ένα αρχείο ουτιπυτ και ανάλογα με τις επιλογές που έχουμε διαβάσει από την `FileReading()` εκτελούμε το αντίστοιχο σετ εντολών και γράφουμε τα αποτελέσματα στο ουτιπυτ αρχείο. Όλοι οι πιθανοί συνδυασμοί που μπορούν να προκύψουν αναλύονται στο παραπάνω διάγραμμα ροής.

Κεφάλαιο 4

Έλεγχος

Στο κεφάλαιο αυτό γίνεται ο έλεγχος καλής λειτουργίας του λογισμικού.

4.1 Μεθοδολογία Ελέγχου

Ο έλεγχος του συστήματος αυτού πραγματοποιήθηκε με τη χρήση του στατιστικού πακέτου Stata . Το πακέτο αυτό περιέχει υλοποιημένες στατιστικές συναρτήσεις που πραγματοποιούν τη διαδικασία της μετα-ανάλυσης. Η διαδικασία πραγματοποιείται με τον χρήστη να εισάγει το σει δεδομένων στο πακέτο και να καλεί τις ανάλογες συναρτήσεις με τα κατάλληλα ορίσματα όπως θα δούμε παρακάτω αναλυτικά, έτσι ώστε να προκύψουν τα επιθυμητά αποτελέσματα.

4.2 Αναλυτική παρουσίαση ελέγχου

Στην ενότητα αυτή παρουσιάζουμε αναλυτικά τον έλεγχο του συστήματος σύμφωνα με τη διαδικασία που περιγράφηκε στην προηγούμενη ενότητα. Αρχικά φορτώνουμε το σει δεδομένων ανά κατηγορία, σχηματίζουμε δηλαδή έξι ομάδες όπου η κάθε μία περιέχει τις τρεις πιθανές μορφές των γονοτύπων (αα, αβ, ββ) για τους ασθενείς και τους υγιείς. Στη συνέχεια, για να συμμορφωθούμε με τις υπάρχουσες συναρτήσεις του Stata, από τις έξι αυτές ομάδες καταλήγουμε σε τέσσερις. Οι τέσσερις αυτές θα αποτελούνται από τον συνολικό αριθμό των Α και των Β αλληλομόρφων σε ασθενείς και υγιείς αντίστοιχα. Κατόπιν λοιπόν της ορθής εισαγωγής και επεξεργασίας του σει δεδομένων στο λογισμικό, πληκτρολογούμε την εντολή

metan a0 a1 b0 b1, or log randomi

όπου **metan** είναι η εντολή που χρησιμοποιούμε για τη μετα-ανάλυση, **a0,a1,b0,b1** οι ομάδες γονιδίων ασθενών και υγιών που κατασκευάσαμε νωρίτερα. Με το **or** ορίζουμε τα αποτελέσματα που θα παραχθούν να είναι σε μορφή Odds Ratio ενώ με το **log** εξασφαλίζουμε ότι θα λάβουμε το νεπέριο αυτού. Το **randomi** εξασφαλίζει ότι πραγματοποιούμε το Μοντέλο Τυχαίων Επιδράσεων με την χρήση της μεθόδου Αντίστροφης Διακύμανσης(Random Effects Model. Αντ' αυτού θα μπορούσαμε να είχαμε χρησιμοποιήσει το όρισμα **fixedi** για να πραγματοποιήσουμε έλεγχο και για το μοντέλο Σταθερών Επιδράσεων Fixed effects Model.

Για τα συνεχή δεδομένα πρέπει να ακολουθήσουμε μία διαφορετική διαδικασία. Αρχικά εισάγουμε εκ νέου το σει δεδομένων, σχηματίζοντας εννέα κατηγορίες που αφορούν ανα τρεις

(λόγω γονοτύπων), τον πληθυσμό, την τυπική απόκλιση και τον σταθμισμένο μέσο των ερευνητών. Οι κατηγορίες αυτές θα μειωθούν σε έξι αθροίζοντας τα δεδομένα και συμπυκνώνοντας τα σε κατηγορίες με βάση το γονίδιο τους (A ή B). Αφού πραγματοποιήσουμε την εισαγωγή των δεδομένων λοιπόν πληκτρολογούμε την εντολή

metan popA popB meanA meanB stddevA stddevB, log or randomi

όπου τα **popA** και **popB** συμβολίζουν τον πληθυσμό όσων φέρουν το κάθε γονίδιο, **meanA** και **meanB** τους σταθμισμένους μέσους του κάθε πληθυσμού και τέλος τα **stddevA** και **stddevB** τις τυπικές αποκλίσεις των δύο πληθυσμών. Τα υπόλοιπα ορίσματα παραμένουν ίδια με εκείνα των διακριτών δεδομένων.

Επί παραδείγματι, παρακάτω μπορούμε να δούμε αναλυτικά τα αποτελέσματα που παράγονται από το Stata εκτελώντας την εντολή που επεξηγήσαμε παραπάνω για τα διακριτά δεδομένα.

```
. metan a0 a1 b0 b1, log or randomi
```

Study	log OR	[95% Conf. Interval]		% Weight
1	-0.227	-0.338	-0.116	20.05
2	-0.541	-0.651	-0.432	20.07
3	0.067	-0.062	0.196	19.84
4	-0.122	-0.239	-0.006	20.00
5	-0.773	-0.886	-0.660	20.03
D+I pooled logOR	-0.320	-0.609	-0.032	100.00

```

Heterogeneity: chi-squared = 124.70 (d.f. = 4) p = 0.000
I-squared (variation in OR attributable to heterogeneity) = 96.8%
Estimate of between-study variance Tau-squared = 0.1048

Test of logOR=0 : z= 2.18 p = 0.030

```

Εικόνα 4.1: Παράδειγμα εκτέλεσης του Stata και τα αποτελέσματα του.

Τα αποτελέσματα αυτά συγκρίνονται με τα αντίστοιχα αποτελέσματα του λογισμικού που υλοποιήσαμε. Όπως γίνεται αντιληπτό στο επακόλουθο στιγμιότυπο, τα αποτελέσματα είναι

```

Stata 16.0: xel
A0: [828.0, 755.0, 717.0, 832.0, 655.0] B0: [5044.0, 4495.0, 4591.0, 4626.0, 4633.0] A1: [667.0, 814.0, 417.0, 571.0, 824.0] B1: [3237.0, 2820.0, 2855.0, 2809.0, 2690.0]
Cochran's Q: 124.7011538497861
Tau squared: 0.10480366589103172
Random Effect Inverse Variance: [9.257730952993057, 9.264942634976759, 9.160484827652162, 9.231132468224477, 9.247993340107356]
Random Effect Weighted Mean: -0.3203579607330139
Random Effect Std. Error for each study: [0.0566835906084135, 0.05594715255751255, 0.0660368218513427, 0.05937511780947238, 0.057687952542543865]
Random Effect Confidence Interval for each study: [-0.3384487722814749, -0.11620989709638484, -0.6511278187398015, -0.4318149807143523, -0.0624685741869242, 0.19639576747033916,
Random Effect Std. Error: 0.14718256073872052
Random Effect Limits: [-0.6088357797809061, -0.03188014168512171]
Random Effect Zeta Value: 2.176602711116814
Phi Value: 0.9552448886363869
P-Value: 0.02951022272322623

```

Εικόνα 4.2: Παράδειγμα εκτέλεσης του λογισμικού μας και τα αποτελέσματα του.

πανομοιότυπα συνεπώς μπορούμε να συμπεράνουμε το γεγονός ότι ο έλεγχος μας πραγματοποιήθηκε με επιτυχία και το λογισμικό μας εκτελείται χωρίς λάθη. Αξίζει να σημειωθεί πως οι έλεγχοι έχουν πραγματοποιηθεί για όλες τις πιθανές παραλλαγές τις οποίες δύναται να αιτηθούν οι χρήστες όπως παραδείγματος χάριν τα διάφορα μοντέλα κληρονομικότητας (Επικρατες, Συνεπικρατές, Υπολειπόμενο) ή τα μοντέλα τυχαίων και σταθερών μεταβλητών.

Παρακάτω παρουσιάζεται ένα δείγμα και του αρχείου που θα λαμβάνει ο χρήστης κατόπιν της εκτέλεσης του προγράμματος.

output.txt - Σημειωματάριο

Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια

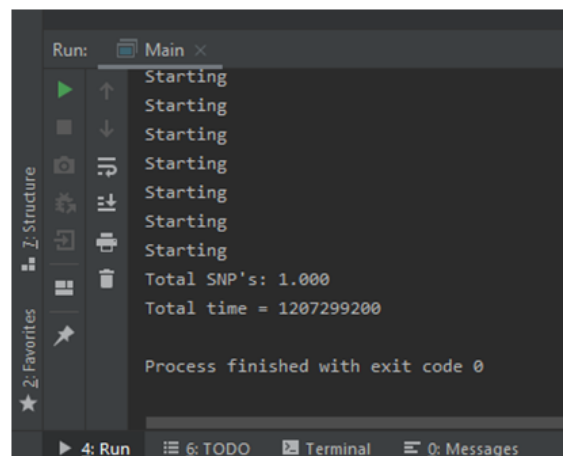
rs	Q	I ²	T ²	Weighted Mean	Cases	Controls	N	StdError	Lower Limit	Upper Limit	Zeta	P-Value	Pos.	Chrom.
rs1	124.7011538497861			96.79233120423379	0.10480366589103172			-0.3203579607330139	2644.0	1757.0	4401.0	0.14718256073872052	-0.6088357797809061	
rs3	61.79674426116611			98.38179177243741	0.20832930001250966			-0.44796149092727455	2644.0	1757.0	4401.0	0.3253889259033558	-1.0857237856978519	
rs2	50.131363064299514			96.01048150748512	0.08481236617252504			-0.23533906681474712	2654.0	1636.0	4290.0	0.1716364230057954	-0.5717464559061061	

Εικόνα 4.3: Αρχείο εξόδου του λογισμικού.

4.3 Χρονικές Αποδόσεις

Η μέθοδος του πολυνηματισμού χρησιμοποιήθηκε για να επιταχυνθεί η διαδικασία της εκτέλεσης του προγράμματος κατά το μέγιστο δυνατό βαθμό. Στο σημείο αυτό θα τρέξουμε το προγράμμα μας για 1000, 100.000 και 1.000.000 πολυμορφισμούς. Για κάθε πολυμορφισμό τα δοκιμαστικά αρχεία θα περιέχουν δύο έρευνες. Αξίζει να σημειωθεί ότι οι δοκιμές έγιναν σε σύστημα σε οκταπύρηνο επεξεργαστή AMD Ryzen 2700X στα 3,7GHz.

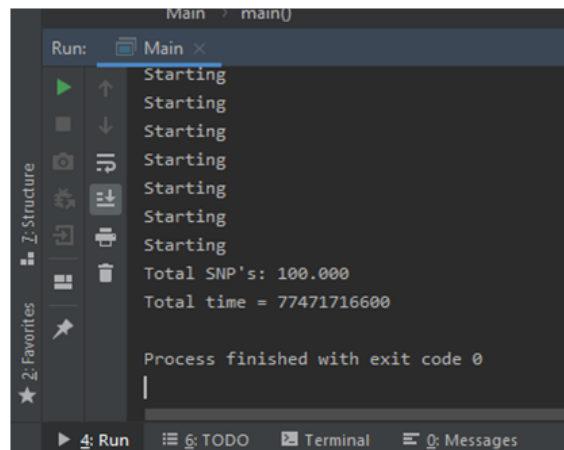
Για 1000 πολυμορφισμούς το πρόγραμμα μας τρέχει σε 0,12072992 second όπως φαίνεται και στην παρακάτω εικόνα.



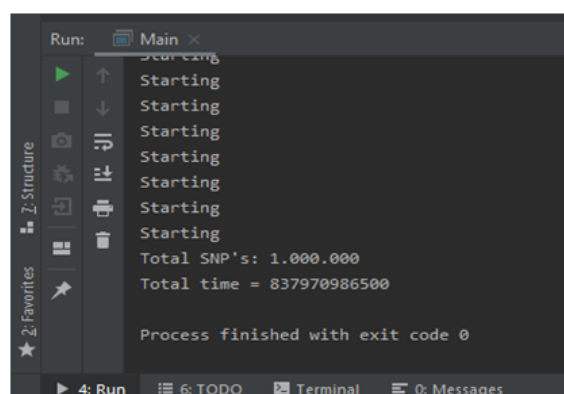
Εικόνα 4.4: Χρόνος εκτέλεσης προγράμματος με 1000 πολυμορφισμούς.

Για 100.000 πολυμορφισμούς όπως είναι λογικό ο χρόνος εκτέλεσης αυξάνεται αισθητά στα 77,4717166 seconds δηλαδή κοντά στο ένα λεπτό και 17 δευτερόλεπτα.

Τέλος στους 1.000.000 πολυμορφισμούς ο χρόνος εκτέλεσης αυξάνεται ραγδαία και φτάνει τα 837,9709865 seconds δηλαδή περίπου στα 14 λεπτά.



Εικόνα 4.5: Χρόνος εκτέλεσης προγράμματος με 100.000 πολυμορφισμούς.



Εικόνα 4.6: Χρόνος εκτέλεσης προγράμματος με 1.000.000 πολυμορφισμούς.

Μέρος **III**

Επίλογος

Κεφάλαιο 5

Συμπεράσματα

Κατα τις έρευνες GWAS εκτελείται μια ευρεία ανίχνευση του DNA σε εκατοντάδες χιλιάδες κοινές παραλλαγές αλληλουχίας σε χιλιάδες ανθρώπους με σκοπό τη χαρτογράφηση σημείων που σχετίζονται με κάποια ασθένεια ή νόσο. Ο όγκος των δεδομένων που πρέπει να επεξεργαστούν είναι τεράστιος για να βγουν τα επι μέρους αποτελέσματα της κάθε έρευνας. [16]

Η μετα-ανάλυση των ερευνών αυτού του τύπου συμβάλλει στην πιο αξιόπιστη εξαγωγή συμπερασμάτων πάνω στο αντικείμενο της έρευνας. Αυτό επιτυγχάνεται μέσω της από κοινού επεξεργασίας των ήδη συλλεγμένων αποτελεσμάτων των επι μέρους ερευνών με διάφορες στατιστικές μεθόδους, οι οποίες εν τέλει θα παράξουν ένα τελικό συμπέρασμα συναρτήσει όλων των προηγούμενων. Η μέθοδος αυτή μας εξασφαλίζει τα καλύτερα δυνατά αποτελέσματα καθώς μας δίνει τη δυνατότητα να χρησιμοποιήσουμε ένα πιο ευρύ φάσμα δεδομένων που θα προέρχεται από όλες τις ήδη υπάρχουσες έρευνες πάνω στο αντικείμενο, μας επιτρέπει να φιλτράρουμε τις έρευνες καθώς και να τις λάβουμε υπόψη μας ανάλογα με τη σημαντικότητα τους.

Όλα τα παραπάνω είναι απαραίτητο να γίνονται στο συντομότερο δυνατό χρόνο, καθώς ο όγκος των δεδομένων όπως επίσης και των εξαγόμενων αποτελεσμάτων είναι τεράστιος. Είναι πρόδηλο λοιπόν πως η εξέλιξη της τεχνολογίας και κατά συνέπεια της πληροφορικής έχει οδηγήσει την ερευνητική κοινότητα στην αναζήτηση εργαλείων που όχι μόνο απλοποιούν αλλά και επιταχύνουν τέτοιες διαδικασίες.

Η συνεισφορά της παρούσας διπλωματικής εργασίας είναι πολύπλευρη. Το πρώτο αφορά τη δημιουργία ενός λογισμικού που θα πραγματοποιεί αξιόπιστα τη μετα-ανάλυση των Μελετών Συσχέτισης Ολόκληρου του Γονιδιώματος και θα παράγει ένα εύρος αποτελεσμάτων για τον εκάστοτε χρήστη, με στόχο την όσο το δυνατόν χρηστικότερη παρουσίαση τους στον ερευνητή.

Το δεύτερο σκέλος αφορά την διασφάλιση πως τα αποτελέσματα που θα παραχθούν από το ανωτέρω λογισμικό, θα παραχθούν με τον ταχύτερο δυνατό ρυθμό. Για το λόγο αυτό έχουν χρησιμοποιηθεί τα threads της JAVA που διαμοιράζουν τον όγκο της εργασίας σε επιμέρους κομμάτια τα οποία επεξεργάζεται ταυτόχρονα για να ελαχιστοποιήσει το χρόνο εκτέλεσης του προγράμματος.

Επιπλέον, το παρόν είναι ένα λογισμικό ανοιχτού κώδικα που σημαίνει πως μπορεί να χρησιμοποιηθεί αλλά και να επεκταθεί από τον οποιονδήποτε. Ταυτόχρονα δε, επιτελεί λειτουργίες που δεν συμπεριλαμβάνονται σε παρεμφερή λογισμικά όπως η δυνατότητα επιλογής

του γενετικού μοντέλου από τον ερευνητή.

Συμπερασματικά, το λογισμικό που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής είναι ένα πλήρες λογισμικό για την εκτέλεση μετα-ανάλυσης Μελετών Συσχέτισης Ολόκληρου του Γονιδιώματος, το οποίο καθιστά την εξαγωγή αποτελεσμάτων αξιόπιστη ενώ ταυτόχρονα πραγματοποιείται σε ανταγωνιστικό χρόνο.

5.1 Μελλοντικές Επεκτάσεις

Το λογισμικό που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής εργασίας θα μπορούσε να βελτιωθεί και να επεκταθεί περαιτέρω, τουλάχιστον ως προς δύο κατευθύνσεις. Συγκεκριμένα, αναφέρονται τα ακόλουθα:

- Ενσωμάτωση του παρόντος λογισμικού σε ένα ευρύτερο λογισμικό που θα περιέχει και τη δυνατότητα πραγματοποίησης Robust μετα-ανάλυσης GWAS δεδομένων εντός του οποίου θα ενσωματωθεί πρωτόκολλο με τη χρήση της τεχνολογίας Blockchain για την ασφαλή μεταφορά των δεδομένων μέσω διαδικτύου και τη διασφαλισή τους .
- Βελτιστοποίηση όσων αφορά τους χρόνους εκτέλεσης του προγράμματος αλλά και βελτιώσεις που αφορούν τη διευκόλυνση της χρήσης του λογισμικού από τον χρήστη αλλά και το πλήθος των δεδομένων που λαμβάνει με πιθανή την προσθήκη και διάφορων ποσοτικών και ποιοτικών διαγραμμάτων που θα βελτιώσουν σε μεγάλο βαθμό την εμπειρία του χρήστη.
- Επέκταση του παρόντος λογισμικού και ενσωμάτωση μεθόδου για ανάλυση μικροστοιχείων.

Βιβλιογραφία

- [1] Gosling, James and Joy, Bill and Steele, Guy and Bracha, Gilad. *The Java Language Specification, Third Edition*. 2005.
- [2] Pawan Murarka, Motahar Reza και Rama Ranjan Panda. *Analysis of Multithreading in Java for Symbolic Computation on Multicore Processors*. 2014.
- [3] Tutorials Point. *Life Cycle Of A Thread*, 2020.
- [4] Bush WS, Moore JH. *Chapter 11: Genome-wide association studies. PLoS computational biology*, 2012.
- [5] Marchini J, Howie B. *Genotype imputation for genome-wide association studies. Nature Reviews Genetics*, 2010.
- [6] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ Sham PC. *PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics*, 2007.
- [7] Zeggini E, Ioannidis JP . *Meta-analysis in genome-wide association studies. Pharmacogenomics*, 2009.
- [8] Pearson TA, Manolio TA . *How to interpret a genome-wide association study. JAMA*, 2008.
- [9] Genetic Alliance; The New York Mid Atlantic Consortiumfor Genetic και Newborn Screening Services. *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals*. 2009.
- [10] Altshuler D Ardlie K Barroso I Boehnke M Cornelis MC Frayling TM Grallert H Grarup N Groop L Hansen T Hattersley AT Hu FB Hveem K Illig T Kuusisto J Laakso M Langenberg C Lyssenko V McCarthy MI Morris A Morris AD Palmer CN Payne F Platou CG Scott LJ Voight BF Wareham NJ Zeggini E Ioannidis JP. Salanti G, Southam L. *Underlying genetic models of inheritance in established type 2 diabetes associations*. 2009.
- [11] Ahn, EunJin, and Hyun Kang. *Introduction to systematic review and meta-analysis. Korean journal of anesthesiology*, 2018.
- [12] Evangelou E, Ioannidis JPA. *Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics*, 2013.

- [13] Fujian Song, Lee Hooper και Loke YK. *Publication bias: What is it? How do we measure it? How do we avoid it?* *Open Access Journal of Clinical Trials*, 2013.
- [14] Philip Sedgwick. *Meta-analyses: what is heterogeneity?* *BMJ*, 2015.
- [15] Matthias Egger, George Davey Smith και Andrew N Phillips. *Meta-analysis: Principles and procedures*. *BMJ*, 315(7121):1533–1537, 1997.
- [16] John Paul SanGiovanni, Richard Rosen και S Kaushal. *Application and Interpretation of Genome-Wide Association (GWA) Studies for Informing Pharmacogenomic Research - Examples from the Field of Age-Related Macular Degeneration*. *Current molecular medicine*, 14, 2014.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
etc	et cetera
GWAS	Genome Wide Association Study
GWA	Genome Wide Association
GWAX	Genome-Wide Association Study by Proxy
DNA	Deoxyribonucleic Acid
ARMD	Age-Related Macular Degeneration
WTCCC	Wellcome Trust Case Control Consortium
CPU	Central Processing Unit

Απόδοση ξενόγλωσσων όρων

Απόδοση

Αναλογία πιθανοτήτων

Εικονικό φάρμακο

Ένδειξη

Δεξαμενή νημάτων

Κεντρική Μονάδα Επεξεργασίας

Μελέτες Συσχέτισης Ολόκληρου Γονιδιώματος

Μεταβλητές υπόστασης

Μετα-Ανάλυση

Μοντέλο των Σταθερών Αποτελεσμάτων

Μοντέλο των Τυχαίων Αποτελεσμάτων

Νήμα

Ομάδα Ασθενών

Πολυμορφισμοί ενός νουκλεοτιδίου

Πολυνηματισμός

Συστηματική Ανασκόπηση

Σφάλμα Δημοσίευσης

Υγιής Ομάδα

Ξενόγλωσσος όρος

Odds ratio

Placebo

Evidence

Thread pool

Central Processing Unit

Genome Wide Association Studies

Instance variables

Meta-Analysis

Fixed-Effects Model

Random-Effects Model

Thread

Cases

Single Nucleotide Polymorphisms

Multithreading

Systematic Review

Publication Bias

Control

