

## Εισαγωγή και μέθοδοι

Μια μελέτη GWAS ορίζεται από το National Institutes of Health ως η μελέτη των κοινών γενετικών παραλλαγών σε ολόκληρο το ανθρώπινο γονιδίωμα που έχει σχεδιαστεί για να εντοπίσει τη γενετική συσχέτιση με παρατηρούμενα χαρακτηριστικά (Pearson, 2008).

Για να πραγματοποιήσουμε τη GWAS μελέτη από το τεχνιτό σετ δεδομένων ακολουθήσαμε τα παρακάτω βήματα με την αξιοποίηση του εκάστοτε εργαλείου που δημιουργήσαμε. Αρχικά, παίρνοντας τα δύο αρχεία δεδομένων, για τα case και control άτομα, παρατηρήσαμε ότι το σύνολο των SNP ήταν 195145 σε κάθε αρχείο αντίστοιχα.

Ως πρώτο βήμα, τρέξαμε το πρόγραμμα μας ώστε να απομονώσουμε τις αλληλικές συχνότητες για τα reference και τα alternative αλληλόμορφα στα cases και τα controls ξεχωριστά όσο και στο ενωμένο αρχείο για κάθε αλληλόμορφο. Για να πραγματοποιηθεί αυτή η ανάλυση, χρησιμοποίησαμε αρχικά τη συνάρτηση genotype\_counts η οποία παίρνει ως όρισμα τις γραμμές από κάθε αρχείο και υπολογίζει τα πλήθη των γονοτύπων για τα ομόζυγα ως προς το reference αλληλόμορφο, για τα ομόζυγα ως προς το alternative αλληλόμορφο και για τα ετερόζυγα άτομα. Χρησιμοποιώντας αυτές τις τιμές στη συνέχεια, τρέξαμε τη συνάρτηση allele\_freq η οποία υπολογίζει την αλληλική συχνότητα για το reference (p) και το alternative (q) αλληλόμορφο διαρώντας το άθροισμα των φορών που παρατηρείται το κάθε αλληλόμορφο σε ένα πληθυσμό με το συνολικό αριθμό των αντιγράφων όλων των αλληλομόρφων στο συγκεκριμένο γενετικό τόπο στον πληθυσμό. Ανάλογα με το αρχείο που βάζουμε να διαβαστεί κάθε φορά, παίρνουμε και τις αντίστοιχες αλληλικές συχνότητες (cases, controls, ενωμένο).

Στη συνέχεια, αφού φιλτράραμε τα αρχεία μας για p και q μεγαλύτερα του 0.05 καταλήξαμε σε αρχεία που περιείχαν 147878 SNPs. Σε αυτά εφαρμόσαμε την HWE για το ενωμένο αρχείο. Η συνάρτηση αυτή υπολογίζει τα p-values της απόκλισης των παρατηρούμενων από τις αναμενόμενες γονοτυπικές συχνότητες στον συνολικό πληθυσμό. Σύμφωνα με την Αρχή των Hardy-Weinberg (ιδανική κατάσταση), οι γονοτυπικές και οι αλληλικές συχνότητες ενός πληθυσμού θα παραμείνουν σταθερές από γενιά σε γενιά υπό την απουσία οποιασδήποτε εξελικτικής επηρροής (όπως είναι η μετάλλαξη, η φυσική επιλογή ή η γενετική παρέκκλιση). Επομένως οι αναμενόμενες συχνότητες των γονοτύπων διατηρούν τις αναλογίες  $p^2$ ,  $q^2$  και  $2pq$ . Ελέγχοντας λοιπόν για πιθανή ισορροπία HW, έχουμε ως αρχική υπόθεση ότι ο πληθυσμός βρίσκεται όντως σε ισορροπία και σύμφωνα με τις p-value που θα προκύψουν από τον έλεγχο, απορρίπτουμε ή όχι την αρχική υπόθεση.

Έπειτα, αφαιρόντας τα p-values που ήταν μικρότερα από 0.001 καταλήξαμε σε αρχείο με 147212 SNPs. Σε αυτά τρέξαμε την συνάρτηση

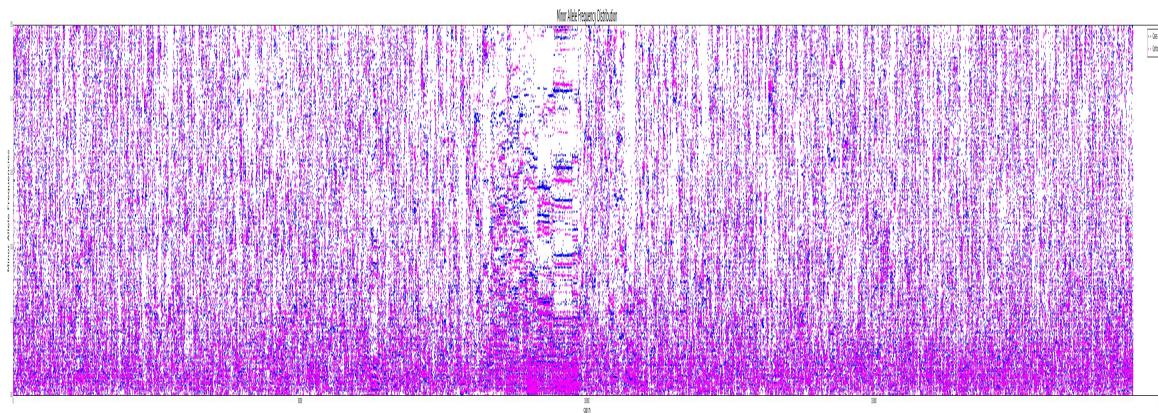
`allelic_association_test` η οποία επιστρέφει τα p-values της απόκλισης μεταξύ των αλληλομόρφων των cases από αυτά των controls για ένα συγκεκριμένο SNP, καθώς και το odds ratio μεταξύ υγιών και ασθενών μαρτύρων. Με τις μελέτες συσχέτισης διάφοροι πολυμορφισμοί μπορούν να εξετασθούν για πιθανή συσχέτιση της παρουσίας τους με ύπαρξη συγκεκριμένου φαινοτύπου (ευαισθησία σε κάποια ασθένεια). Κατα τον έλεγχο της αρχικής υπόθεσης όπου δεν υφίσταται σύνδεση μεταξύ γονοτύπου και φαινοτύπου, χαμηλές τιμές p-value τείνουν να απορρίψουν την αρχική υπόθεση και άρα θεωρούμε πως υπάρχει συσχέτιση. Επιλέγοντας το όριο  $10^{-6}$  για τα p-values του association test βρίσκουμε 75 στατιστικά σημαντικά SNPs (Clark et al. 2011).

Επόμενο βήμα στην ανάλυση μας είναι η εξεταση των προτύπων σύνδεσης στις περιοχές των statistical significant snps που εχουν προκυψει απο την μελέτη σύνδεσης. Η ανισορροπία σύνδεσης (Linkage disequilibrium, LD) είναι η μη τυχαία συσχέτιση των αλληλομόρφων που βρίσκονται σε κοντινές θέσεις πάνω στο γονιδίωμα (Doris, 2002). Όταν ένα συγκεκριμένο αλληλόμορφο σε ένα γενετικό τόπο συγκληρούμεται μαζί με κάποιο άλλο αλληλόμορφο που βρίσκεται σε ένα δεύτερο γενετικό τόπο πάνω στο ίδιο χρωμόσωμα, τότε οι τόποι βρίσκονται σε ανισορροπία σύνδεσης. Έστω δύο γειτονικοί γενετικοί τόποι A και B, με δύο αλληλόμορφα: A, α και B, β αντίστοιχα. Απλότυπος ονομάζεται η διευθέτηση των αλληλομόρφων πάνω στο ίδιο χρωμόσωμα. Επομένως συμβολίσουμε με PAB την παρατηρούμενη συχνότητα του απλοτύπου που αποτελείται από αλληλόμορφα A και B. Καθώς τα δεδομένα που διαθέτουμε μας παρεχουν μόνο γονοτυπικές συχνότητες για τις διάφορες θέσεις, είναι αδύνατο να γνωρίζουμε την διευθέτηση των αλληλομόρφων A και B στο χρωμόσωμα του κάθε ατόμου. Για να εκτιμήσουμε τις παρατηρούμενες απλοτυπικές συχνότητες υλοποιούμε τον αλγόρυθμο 'Expectation\_maximization' (EM) στην ομόνυμη συνάρτηση. Η αναμενόμενη συχνότητα του απλοτύπου υποθέτοντας την ανεξάρτητη διάταξη των αλληλόμορφων στις δύο θέσεις, υπολογίζεται από το γινόμενο ( $PA \times PB$ ) όπου PA είναι η συχνότητα του αλληλόμορφου A και PB είναι η συχνότητα του αλληλομόρφου B. Έτσι καταλήγουμε στον υπολογισμό των σύνηθες στατιστικών μέτρων για την ποσοτικοποίηση της ανισορροπίας σύνδεσης ανάμεσα στις δυο θέσεις. Ο έλεγχος την σύνδεσης των ενοχοποιητικών SNPs που προσδιορίσαμε, με άλλους γειτονικούς δείκτες πραγματοποιήθηκε κατά ζεύγη στο φιλτραρισμένο αρχείο όπου έχουν αφερεθεί τα SNP με  $MAF < 0.05$  και  $HWE\text{value} < 0.01$ . Εξετάζεται η τιμή της παραμέτρου D' και θελουμε οι συνδεδεμένες μας θέσεις να έχουν τιμή μεγαλύτερη από 0.9. Ο έλεγχος επεκτείνεται και προς τις δυο κατευθύνσεις εκατέροθεν της θέσης στόχου και είναι σχεδιασμένος να σταματήσει 1000 θέσεις από την τελευταία συνδεδεμένη θέση που εντπόπισε, καθώς όσο απομακρύνεται η σύνδεση, όπως είναι αναμενόμενο, χάνεται.

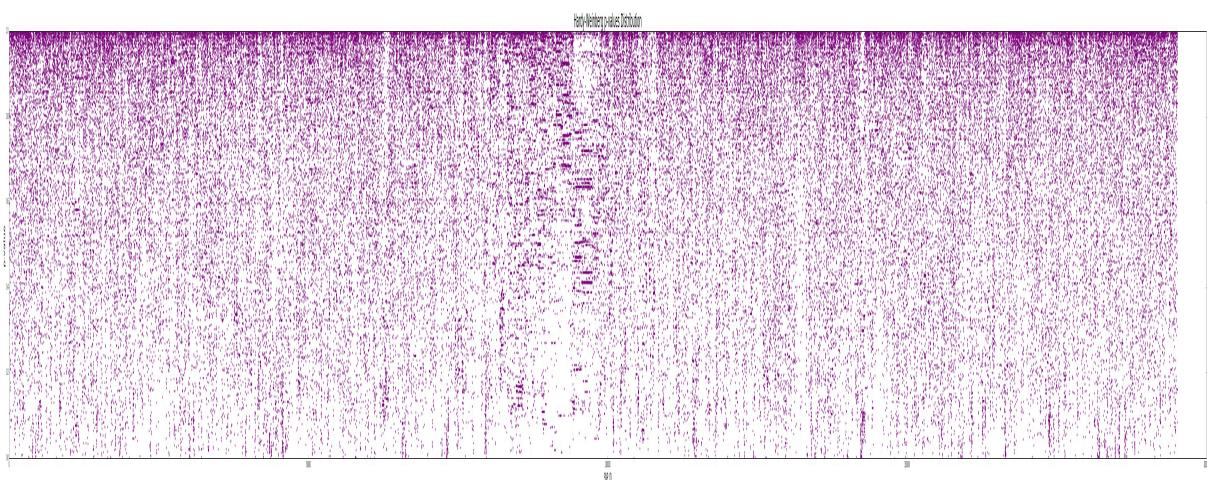
Τελευταίο βήμα είναι η άντληση πληροφοριών απο την βάση δεδομένων Ensembl για τα SNPs ενδιαφέροντος. Για κάθε ενα από αυτά πραγματοποιείται αίτημα τύπου get από την πλατφόρμα Variant Effect Predictor της βάσης δεδομένων, όπου περιέχει πληροφορίες σχετικά με την επίδραση του δείκτη σε άλλα γονίδια, μετάγραφα και ρυθμιστικές περιοχές. Το αίτημα πραγματοποιείται με βάση την θέση του πολυμορφισμού και οι συντεταγμένες του πρέπει να είναι σύμφωνοι την έκδοση GRCh38/hg38 της συναρμολόγησης του ανθρώπινου

γονιδιώματος. Καθώς όμως οι συντεταγμένες μας είναι με βάση το GRCh36/hg18 πρέπει πρώτα να γινει η μετατροπή στην πιο πρόσφατη έκδοση.

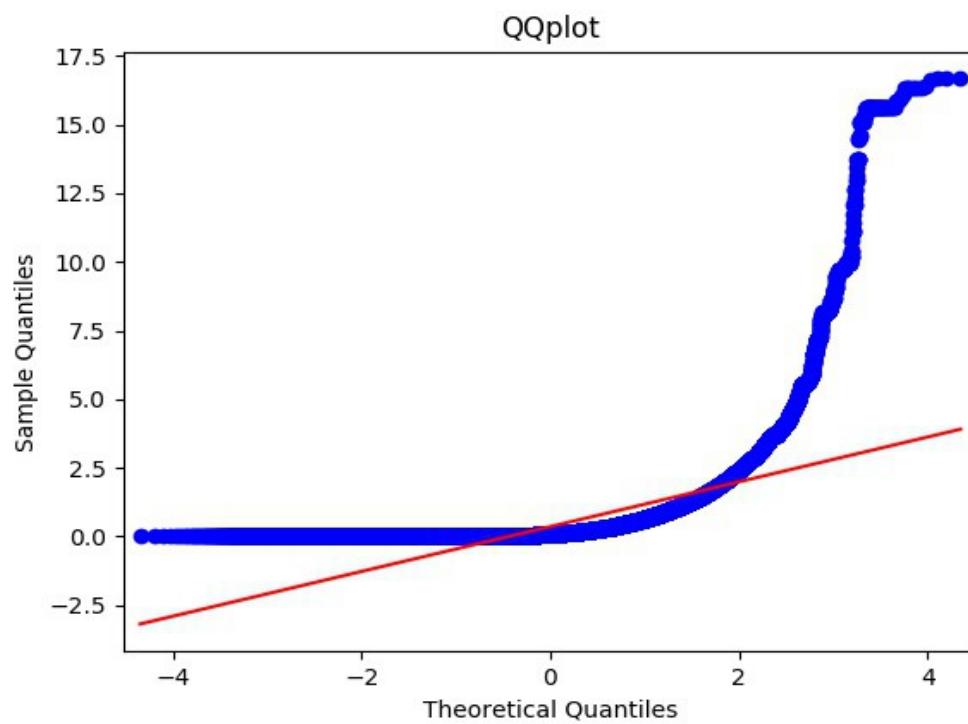
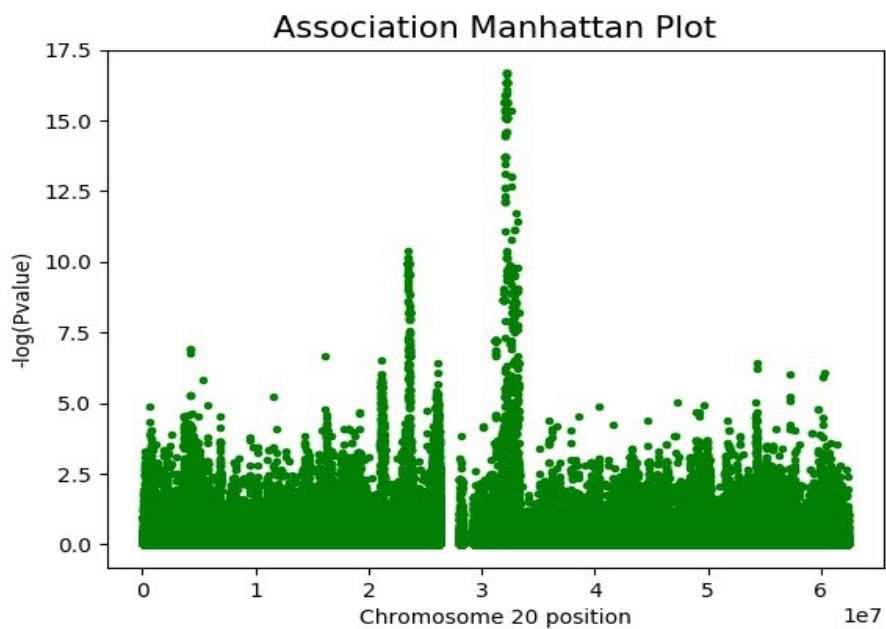
## Αποτελέσματα



Εικόνα 1: MAFS :Υπάρχει σε καλή ανάλυση στο github: <https://github.com/marivasi5/projectara>



Εικόνα 2: HWE:Υπάρχει σε καλή ανάλυση στο github: <https://github.com/marivasi5/projectara>



## Συμπεράσματα

Μείνετε συντονισμένοι για την παρουσίαση μας.

Βιβλιογραφία:

1. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. *Nature protocols*. 2011.
2. Doris, P. A. (2002). Hypertension genetics, single nucleotide polymorphisms, and the common disease: common variant hypothesis. *Hypertension*, 39(2), 323-331.
3. Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama*, 299(11), 1335-1344. 15 Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1), 7-24.