ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΟΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Μεθοδολογίες μελέτης απλότυπων σε μελέτες γενετικής συσχέτισης

Λούκας Αλέξιος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέποντες

Μπάγκος Παντελής, Επίκουρος Καθηγητής

Ιωαννίδης Αναστάσιος, Επισκέπτης Επίκουρος Καθηγητής

Λαμία, 2010

Εξεταστική επιτροπή

1. Μπάγκος Παντελής (επιβλέπων)
2. Αδάμ Μαρία
3. Ιωαννίδης Αναστάσιος (επιβλέπων)

ΕΥΧΑΡΙΣΤΙΕΣ

Η εργασία αυτή έλαβε χώρα στην Λαμία, στο Πανεπιστήμιο Στερεάς Ελλάδος, στο Τμήμα Πληροφορικής με εφαρμογές στην βιοϊατρική. Η ολοκλήρωσή της οφείλεται όχι μόνο στον συγγραφέα, αλλά και στους ανθρώπους που ακολουθούν.

Θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές μου κ. Μπάγκο Παντελή και κ. Ιωαννίδη Αναστάσιο που με επέλεξαν για την εκπόνηση της εργασίας αυτής, καθώς και για τις συμβουλές τους πάνω στο θέμα και το κείμενό της. Επίσης θα ήθελα να ευχαριστήσω τους φίλους, τους συμφοιτητές, την οικογένειά μου, τους γονείς μου αλλά και τον Σεβασμιώτατο Μητροπολίτη Ηλείας και Ωλένης Γερμανό Παρασκευόπουλο για την υποστήριξη που μου παρείχαν κατά την διάρκεια των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Η απλοτυπική ανάλυση αποτελεί πλέον αναπόσπαστο κομμάτι της γονιδιωματικής ανάλυσης. Ο απλότυπος είναι ένα σύνολο πολυμορφισμών που παρουσιάζονται στο DNA και είναι στατιστικά συνδεδεμένοι μεταξύ τους λόγω γειτνίασης. Από τις κλασικές μεθόδους αλληλούχισης, η γνώση του γονότυπου δεν μας δίνει μονοσήμαντα την πληροφορία για τους απλότυπους και για αυτό μας χρειάζονται αλγόριθμοι στατιστικής φύσεως (το πρόβλημα της φάσης των απλότυπων). Στις μελέτες γενετικής συσχέτισης χρησιμοποιούνται ολοένα και περισσότερο οι απλότυποι για την διερεύνηση της συσχέτισης κοινών πολυμορφισμών με ασθένειες. Η δημιουργία του «χάρτη» του ανθρώπινου γονιδιώματος (HapMap) ανοίγει νέες πόρτες με την μεγάλη αυτή βάση δεδομένων και ενθαρρύνει την επιστημονική κοινότητα για περεταίρω ενασχόληση. Σε αυτήν εδώ την εργασία συνοψίζονται οι μέθοδοι που χρησιμοποιούνται σε μελέτες γενετικής συσχέτισης με απλότυπους, αναλύονται τα μαθηματικά μοντέλα για αυτές τις μεθόδους και παρουσιάζονται αναλυτικά παραδείγματα με την χρήση των λογισμικών Stata και Haploview. Η εφαρμογή των μεθόδων γίνεται σε πραγματικά δεδομένα τα οποία πήραμε από δημοσιευμένες μελέτες.

Λέξεις–Κλειδιά: Απλότυπος, Συσχέτιση, Μονοσημειακός Πολυμορφισμός, HapMap, Ασθενείς-Μάρτυρες, Ανάλυση

ABSTRACT

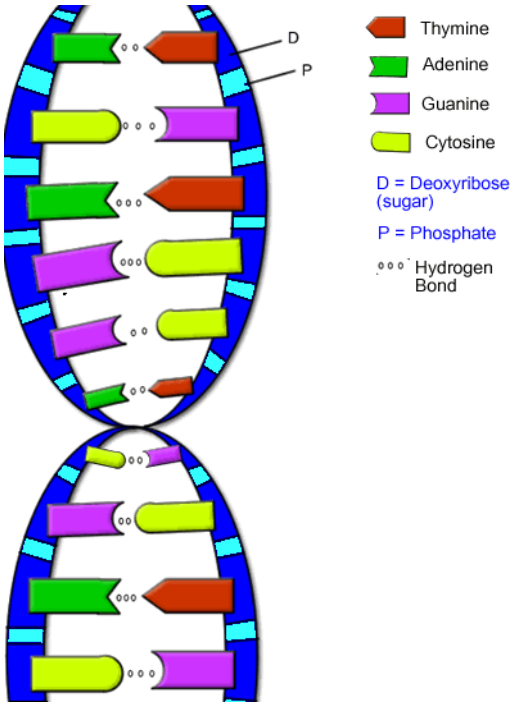
Haplotype analysis constitutes an integral part of genomic analysis. A Haplotype is referred to as a set of single nucleotide polymorphisms that we see in certain DNA regions, and they are statistically associated. From the classic sequencing methods, the knowledge alone of the genotype, doen't give us all the information we need about haplotypes, that is why we need to use statistical inference algorithms (haplotype inference problem). In genetic association studies more and more we see haplotypes used for association between single nucleotide polymorphisms and diseases. The creation of the human genome "map" known as HapMap opens new doors with this large database and encourages the scientific community to further activity. In this thesis we summarize the methods that are used in genetical association with haplotypes and we analyze the mathematical models that these methods use. We present analytical examples with the use of Stata and Haploview software. These methods are applied in real data, taken from published studies.

Key-Words: Haplotype, Association, Single Nucleotide Polymorphism, HapMap, Case-Control, Analysis

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή στους απλότυπους

Η λέξη απλότυπος προέρχεται από τα παράγωγα «απλοειδής» και «γονότυπος». Ουσιαστικά είναι ένα σύνολο πολυμορφισμών σε διάφορα σημεία του DNA (εικόνα 1), οι οποίοι είναι στατιστικά συνδεδεμένοι μεταξύ τους, δηλαδή η εμφάνιση του ενός κάνει πολύ πιθανή και την εμφάνιση άλλων. Μεταδίδονται μαζί σαν ένα ενιαίο κομμάτι πάνω στο γενετικό υλικό (HapMap 2005).



Εικόνα 1: Η δομή του DNA

Έχοντας στην κατοχή την γνώση κάποιων απλότυπων για το γονιδίωμα του ανθρώπου, ίσως έχουμε στα χέρια μας πολύτιμες πληροφορίες που θα μας βοηθήσουν να διερευνήσουμε σε μεγαλύτερο βαθμό την γενετική συμπεριφορά κάποιων ασθενειών. Ήδη γίνεται μια παγκόσμια προσπάθεια για την καταγραφή ενός «χάρτη» απλότυπων του ανθρώπινου γονιδιώματος, ο οποίος θα περιγράφει κάποια κοινά πρότυπα στην γενετική διαφοροποίηση μεταξύ των ανθρώπων.

1.2 To International HapMap project

Ο χάρτης που αναφέρεται παραπάνω είναι ελεύθερα προσβάσιμος στους ερευνητές, ώστε να τους βοηθήσει να συσχετίσουν εκείνες τις γενετικές πληροφορίες που επηρεάζουν την υγεία, συμβάλουν στις ασθένειες, στην αλληλεπίδραση των οργανισμών με τα φάρμακα, αλλά και το ρόλο που παίζουν κάποιοι περιβαλλοντικοί παράγοντες. Μια τόσο μεγάλη μελέτη βέβαια διέπεται από κάποιους κανόνες μεθοδολογίας και ηθικής που δημοσιεύτηκαν από την ομάδα των ερευνητών (HapMap 2004).

Η έρευνα σήμερα στρέφεται κυρίως στους λεγόμενους «μονοσημειακούς πολυμορφισμούς» (Single nucleotide polymorphism ή SNP), οι οποίοι δηλώνουν μια σημειακή αλλαγή στο γενετικό υλικό. Υπάρχουν περίπου δέκα εκατομμύρια πολυμορφισμοί που παρατηρούνται στο ανθρώπινο γονιδίωμα. Από αυτούς, μια πολύ μικρή ομάδα είναι άμεσα συνδεδεμένη με ασθένειες, παρενέργειες κλπ.

Σκοπός αυτού του «χάρτη» είναι πρώτον η διεύρυνση αυτής της μικρής ομάδας, και δεύτερον να κάνει προσβάσιμη την γενετική πληροφορία στους ερευνητές, οι οποίοι δεν θα χρειάζεται πλέον να ψάχνουν για πολυμορφισμούς (HapMap 2003).

Σημαντική συμμετοχή σε αυτήν την προσπάθεια υπάρχει και από έναν Έλληνα επιστήμονα, τον Δρ. Πάνο Δελούκα ο οποίος μαζί με συναδέλφους του στο Sanger Center της Μ. Βρετανίας υλοποίησαν το HapMap project (εικόνα 2) (http://www.sanger.ac.uk/research/faculty/pdeloukas/). Όπως εξηγεί ο κύριος Δελούκας με δηλώσεις του *«το διεθνές πρόγραμμα HapMap δεν είναι παρά ένα εργαλείο το οποίο θα βοηθήσει στην αξιοποίηση της πληροφορίας που έχει προκύψει από την ανάγνωση του ανθρώπινου γονιδιώματος. Στην πράξη, θα δίνει το στίγμα κάποιων πολυμορφισμών με ιδιαίτερο ενδιαφέρον πάνω στο γονιδίωμα»*.

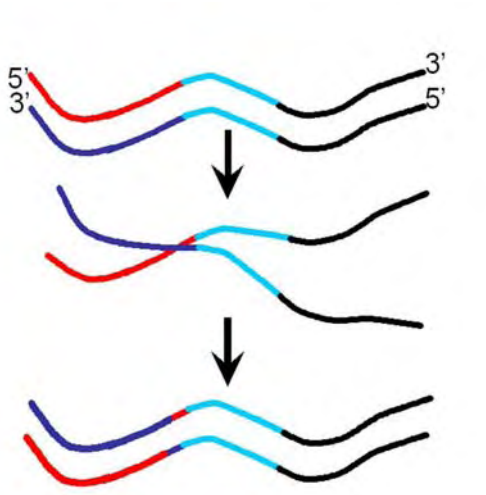
Εικόνα 2: Ο Δρ. Πάνος Δελούκας

Αυτός ο «χάρτης» θα χρησιμοποιεί τους πολυμορφισμούς ως σημεία αναφοράς πάνω στο γονιδίωμα, και ενδέχεται να επιταχύνει τον εντοπισμό γονιδίων που σχετίζονται με ασθένειες και ειδικότερα με ασθένειες όπου η περιβαλλοντική συνιστώσα παίζει καθοριστικό ρόλο στην έκφρασή τους. *«Τα γονίδια των*

μονογονιδιακών ασθενειών, αυτών δηλαδή που η αιτιολογία τους οφείλεται σε μεταλλάξεις ενός και μόνο γονιδίου, δεν αποτελούν αντικείμενο των δικών μας αναζητήσεων. Οι επιστήμονες έχουν αποτελεσματικούς τρόπους να τα εντοπίζουν. Αυτό που μας ενδιαφέρει είναι γονίδια τα οποία κάνουν κάποιους από μας πιο ευαίσθητους σε λοιμώξεις ή αυξάνουν τον κίνδυνο που διατρέχουμε να αναπτύξουμε διαβήτη ή άσθμα ή υπέρταση αν εκτεθούμε σε συγκεκριμένες συνθήκες. Σε τέτοιου είδους αναζητήσεις θα είναι πολύτιμο το HapMap. Για παράδειγμα, αν υποτεθεί ότι θα θέλαμε να αναζητήσουμε τη γενετική διαφορά που καθιστά κάποιους ανθρώπους πιο ευαίσθητους από άλλους στην εμφάνιση διαβήτη. Χωρίς το HapMap θα έπρεπε να πάρουμε δύο ομάδες ανθρώπων, ας πούμε 1.000 ασθενείς με διαβήτη και 1.000 υγιείς, και να εξετάσουμε ποιος από τους 10 εκατομμύρια SNP's είναι διαφορετικός στην ομάδα των ασθενών. Αντιλαμβάνεστε πόσο χρονοβόρα αλλά και ακριβή θα ήταν μια τέτοια διαδικασία, ειδικά καθώς θα έπρεπε να επαναληφθεί για όλες τις ασθένειες που μας ενδιαφέρουν...» μας λέει ο κ. Δελούκας (Deloukas and Bentley 2004).

Σημαντικό στοιχείο στο θέμα του χάρτη είναι η γνωστή διαδικασία του ανασυνδυασμού του DNA (εικόνα 3), που ως γνωστό λαμβάνει χώρα κατά την διάρκεια της μείωσης των γαμετικών κυττάρων. Αυτό έχει ως αποτέλεσμα το γεγονός ότι ο καθένας μας έχει στα γαμετικά του κύτταρα ένα νέο σύνολο γενετικού υλικού, και όχι απλά ένα μείγμα από το DNA του πατέρα και της μητέρας μας. Ο ανασυνδυασμός έχει ιδιαιτερότητες οι οποίες αξιοποιούνται για τη δημιουργία του HapMap (HapMap 2003).

Σύμφωνα με τον κ. Δελούκα *«τα σημεία πάνω στα χρωμοσώματα που συμβαίνει ο ανασυνδυασμός δεν είναι τυχαία. Επίσης οι άνθρωποι είμαστε σχετικά νέο είδος, πράγμα που σημαίνει ότι δεν έχουν γίνει τόσο πολλοί ανασυνδυασμοί κατά τη διάρκεια των αιώνων. Αυτό που αναζητούμε και χαρτογραφούμε εμείς είναι κομμάτια DNA τα οποία έχουν περάσει αυτούσια. Πάνω σε αυτά τα κομμάτια υπάρχουν SNP's οι οποίοι, όταν ολοκληρωθεί το HapMap, θα λειτουργούν ως δείκτες για να καθοδηγούν τους ερευνητές στα σημεία του γονιδιώματος που πρέπει να ψάξουν προκειμένου να βρουν αυτό που τους ενδιαφέρει»*.



Εικόνα 3: Συνοπτικά ο ανασυνδιασμός του DNA

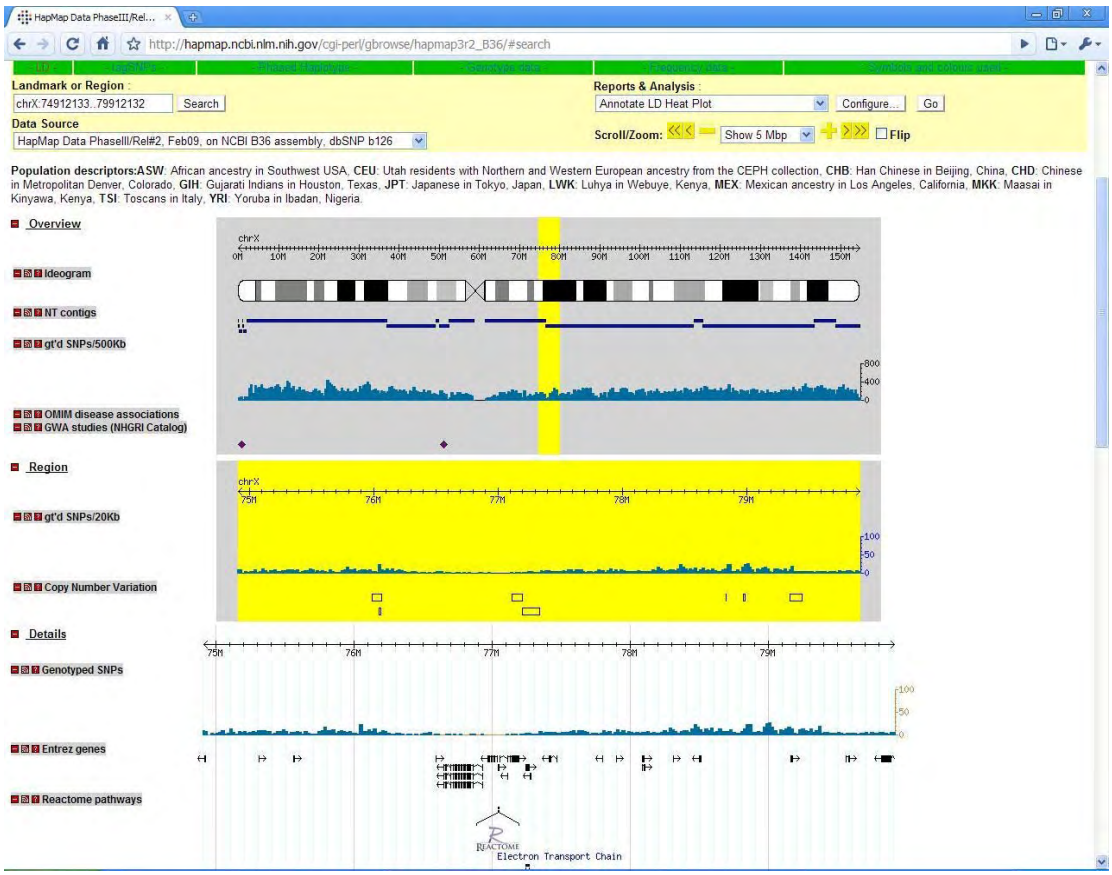
Πρακτικά, αυτό που θα κάνουν οι ερευνητές όταν τελειώσει η δημιουργία του HapMap, θα είναι να αναζητούν αν οι άνθρωποι που φέρουν κάποια νόσο, φέρουν επίσης και κάποιους χαρακτηριστικούς πολυμορφισμούς. Αυτοί οι πολυμορφισμοί θα τους δείξουν σε ποιο σημείο πάνω στα χρωμοσώματα πρέπει να αναζητήσουν το γονίδιο που σχετίζεται με αυτή τη νόσο.

Ο οργανισμός ο οποίος ξεκίνησε την κατασκευή αυτού του χάρτη ονομάζεται «International HapMap Project» (HapMap 2003) και ουσιαστικά είναι μια συνεργασία μεταξύ διαφόρων ερευνητών ακαδημαϊκών κέντρων, μη-κερδοσκοπικών βιοϊατρικών οργανώσεων και ιδιωτικών εταιριών σε Καναδά, Κίνα, Ιαπωνία, Νιγηρία, Ηνωμένο βασίλειο και Ηνωμένες πολιτείες Αμερικής (HapMap 2005). Το λογότυπο του οργανισμού φαίνεται στην εικόνα 6.

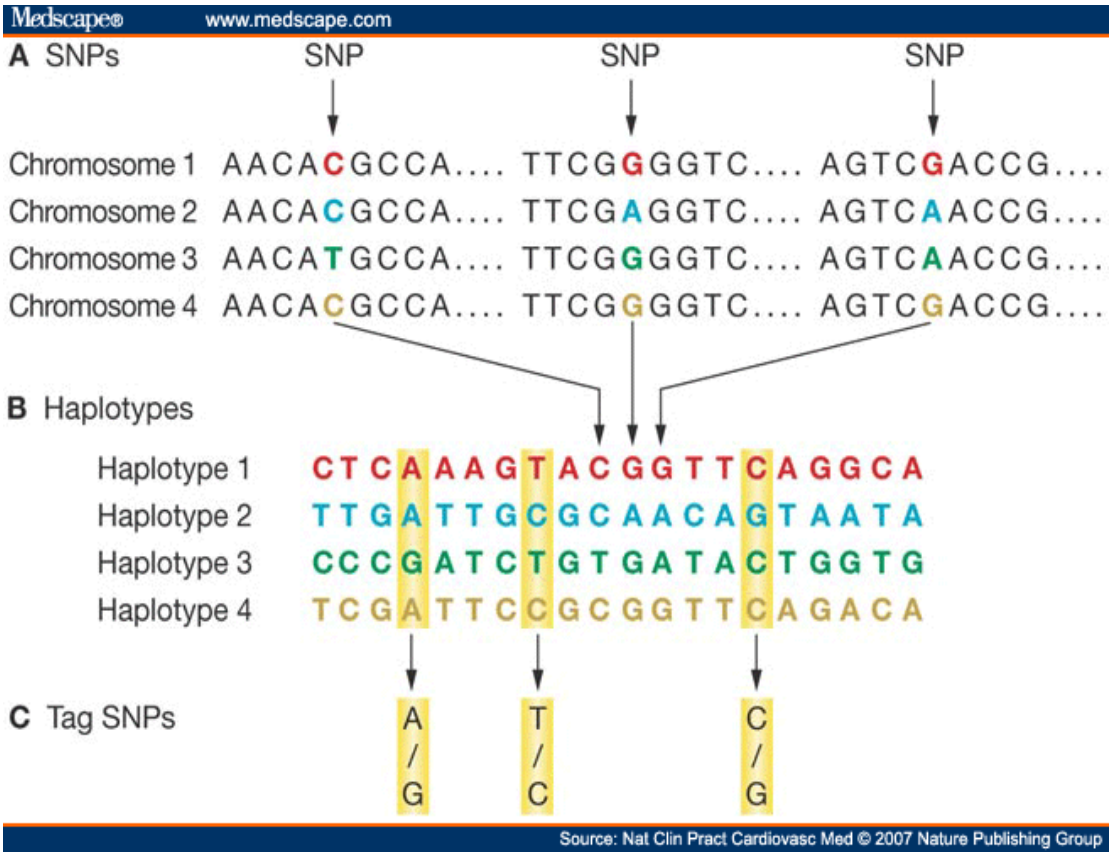
Το project ξεκίνησε τον Οκτώβριο του 2002 και αρχικά εκτιμήθηκε ότι θα κρατούσε τρία χρόνια, ωστόσο βλέπουμε ότι συνεχίζεται ακόμη καθώς τον Μάιο του 2010 είχαμε ακόμη μια δημοσίευση γενετικών δεδομένων (HapMap3 Public Release #3). Οι ενδιαφερόμενοι ερευνητές μπορούν να βρίσκουν τα αποτελέσματα της χαρτογράφησης αυτής στον δικτυακό τόπο http://www.hapmap.org/.

Η ιστοσελίδα επιτρέπει την περιήγηση σε όλες τις περιοχές του ανθρώπινου γονιδιώματος, προσφέροντας όλα τα καταγεγραμμένα στοιχεία όπως γνωστοί απλότυποι, πολυμορφισμοί, γονίδια, εσώνια-εξώνια κλπ. Σαν παράδειγμα στην εικόνα 4 που ακολουθεί, έχουμε επιλέξει μια περιοχή του χρωμοσώματος X προς επεξεργασία μέσω της εφαρμογής που παρέχεται από την ιστοσελίδα του HapMap. Η εφαρμογή μας δείχνει όλα τα δεδομένα που έχει καταχωρημένα και χωρίζει την

περιοχή του γονιδιώματος που ερευνούμε σε τρία επίπεδα μεγέθυνσης. Φυσικά μπορούμε να μεγενθύνουμε επιπλέον την εικόνα και να δούμε από πιο κοντά τις βάσεις. Όσο μεγεθύνουμε και μπαίνουμε σε λεπτομέρειες, τόσο περισσότερο φαίνονται οι καταχωρημένες πληροφορίες των πολυμορφισμών. Ο χρήστης μπορεί να διαλέξει και να εξειδικεύσει την αναζήτησή του με μια πληθώρα επιλογών, σε μια μεγάλη βάση βιολογικών δεδομένων, η οποία είναι εξαιρετικά οργανωμένη. Το μεγαλύτερο πλεονέκτημα είναι ότι είναι εντελώς ελεύθερη η προσπέλαση σε αυτήν, καθώς επίσης και ότι βρίσκεται φορτωμένη στο διαδίκτυο, πράγμα που την κάνει εύκολα προσβάσιμη από οποιοδήποτε μέρος.



Εικόνα 4: Το χρωμόσωμα Χ στον χάρτη του HapMap



Εικόνα 5: Πως δημιουργούνται οι απλότυποι από τους πολυμορφισμούς

1.3 Μονονουκλεοτιδικός πολυμορφισμός

Παρόλο που δυο οποιοιδήποτε άνθρωποι έχουν κοινό το 99.5% του γονιδιώματός τους (HapMap 2003), κάποιοι σε μια συγκεκριμένη θέση στο γενετικό τους υλικό μπορεί να έχουν την βάση A (αδενίνη) ενώ κάποιοι άλλοι την G (γουανίνη). Μια τέτοια αλλαγή (A->G) είναι γνωστή με το όνομα μονονουκλεοτιδικός πολυμορφισμός (single-nucleotide polymorphism ή SNP) και μπορεί να συμβεί σε διάφορα μέρη του γονιδιώματος κατά την διάρκεια ζωής ενός οργανισμού. Κάθε περίπτωση (A ή G) ονομάζεται αλληλόμορφο. Μια απεικόνιση των αλληλόμορφων φαίνεται στην εικόνα 7, όπου η πλειοψηφία των ατόμων έχουν την βάση C (κυτοσίνη) ενώ εμφανίζεται σπάνια η T (θυμίνη).

Το HapMap ασχολείται μόνο με κοινούς πολυμορφισμούς, δηλαδή με αυτούς όπου κάθε αλληλόμορφο υπάρχει τουλάχιστον στο 1% του πληθυσμού. Κάθε άνθρωπος έχει δυο αντίγραφα κάθε χρωμοσώματος, εκτός από τα χρωμοσώματα που καθορίζουν το φύλο. Για κάθε SNP, ο συνδυασμός των αλληλόμορφων που έχει ένα άτομο, ονομάζεται γονότυπος. Έτσι υπάρχει η διαδικασία εύρεσης του γονότυπου

ενός ατόμου, σε μια συγκεκριμένη περιοχή του γονιδιώματός του. Μια τέτοια διαδικασία γίνεται με δυο τρόπους. Σε εργαστήρια με την μέθοδο της ηλεκτροφόρησης, αλλά και με στατιστικές μέθοδους όπου χρησιμοποιούμε βέβαια την ήδη υπάρχουσα πληροφορία που έχουμε από το εργαστήριο. Σαν αποτέλεσμα βρίσκουμε όλους τους πολυμορφισμούς και τα αλληλόμορφά τους για μια συγκεκριμένη περιοχή του DNA. Οι πολυμορφισμοί που βρίσκονται σε κοντινή θέση μεταξύ τους και πάνω στο ίδιο χρωμόσωμα, είναι συσχετισμένοι. Αυτό σημαίνει πως αν το αλληλόμορφο από έναν πολυμορφισμό είναι γνωστό, τότε τα αλληλόμορφα των διπλανών πολυμορφισμών, ίσως να μπορούμε να τα προβλέψουμε (HapMap 2003).

Εικόνα 6: Το HapMap project και οι χώρες που το αποτελούν

Αυτό συμβαίνει επειδή κάθε πολυμορφισμός, εμφανίστηκε εξελικτικά ως μια σημειακή μετάλλαξη, η οποία στη συνέχεια πέρασε στους απογόνους περιβαλλόμενη από άλλες, πιο παλιές σημειακές μεταλλάξεις. Συνήθως βέβαια, πολυμορφισμοί που χωρίζονται από μεγάλη απόσταση, δεν εμφανίζουν συσχέτιση. Κάτι τέτοιο συμβαίνει λόγω του ότι χωρίζονται με την διαδικασία ανασυνδιασμού του DNA σε κάθε γενιά, με αποτέλεσμα το ανακάτεμα των ακολουθιών των αλληλόμορφων των δυο χρωμοσωμάτων. Ωστόσο η πληροφορία που ψάχνουμε είναι μια ακολουθία αλληλόμορφων πάνω σε ένα συγκεκριμένο χρωμόσωμα, η οποία δεν έχει διαχωριστεί με τον ανασυνδιασμό του DNA. Μια τέτοια ακολουθία ονομάζεται απλότυπος. Επίσης απλότυπος θεωρείται και ένα σύνολο από μονοσημειακούς πολυμορφισμούς, όχι απαραίτητα σε ακολουθία, αλλά σε μικρή απόσταση μεταξύ τους σε μια περιοχή του DNA (HapMap 2003).

Για να βρούμε τους γενετικούς παράγοντες που εμπλέκονται σε μια ασθένεια, μπορούμε να ακολουθήσουμε την παρακάτω διαδικασία. Πρώτα πρέπει να βρεθεί μια περιοχή ενδιαφέροντος πάνω στο γονιδίωμα (π.χ. περιοχή γονιδίου), πιθανότατα γνωστή από προηγούμενες μελέτες. Έπειτα εντοπίζουμε σε αυτή την περιοχή ένα

σύνολο από πολυμορφισμούς οι οποίοι έχουν ισχυρή συσχέτιση μεταξύ τους. Το σύνολο των αλληλόμορφων από αυτούς τους πολυμορφισμούς θα μας δώσει τον απλότυπο αυτού του ατόμου. Η διαδικασία αυτή συνοψίζεται στην εικόνα 5. Στη συνέχεια βρίσκουμε τον απλότυπο αυτών των πολυμορφισμών σε ένα σύνολο ατόμων, κάποιων που να έχουν την ασθένεια, και κάποιων που να μην την έχουν. Με διάφορες συγκρίσεις μεταξύ των δυο αυτών ομάδων (ασθενείς – μη ασθενείς) μπορούμε να καταλάβουμε ποιες περιοχές και ποιοι απλότυποι παίζουν κάποιο ρόλο στην εμφάνιση και την συμπεριφορά της ασθένειας (Fallin, Cohen et al. 2001).

Εικόνα 7: Μια διαφορετική απεικόνιση του μονοσημειακού πολυμορφισμού

1.4 Το πρόβλημα της εύρεσης του απλότυπου

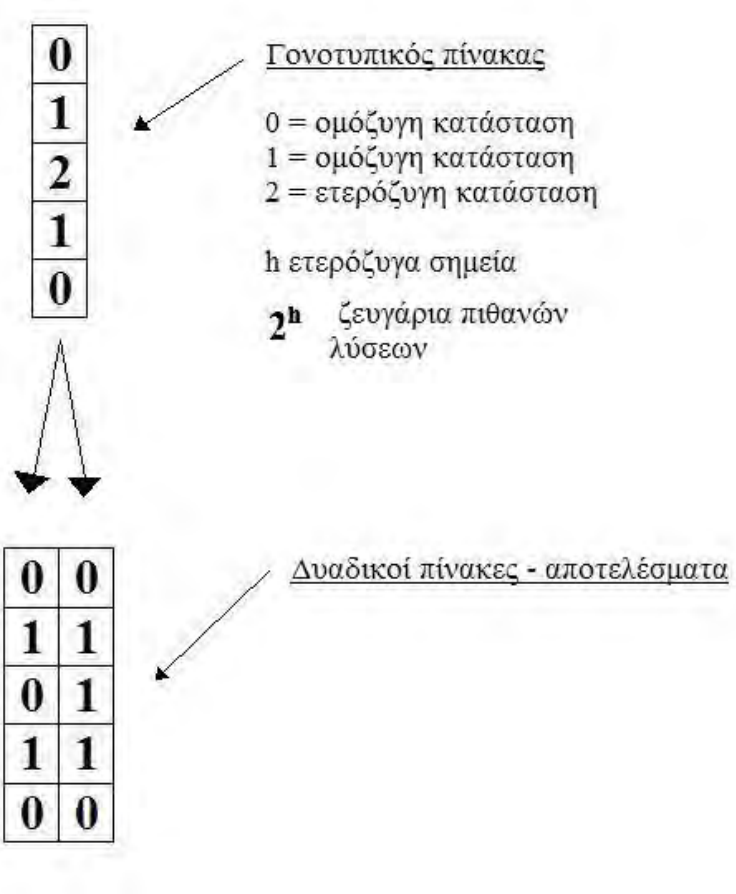
Αντίθετα από τις σπάνιες κληρονομικές ασθένειες, συνδυασμοί διαφόρων γονιδίων αλλά και το περιβάλλον, παίζουν κάποιο ρόλο στην εμφάνιση πιο κοινών παθήσεων (όπως διαβήτης, καρκίνος, παθήσεις της καρδιάς, εγκεφαλικό, κατάθλιψη και άσθμα) αλλά και παρενεργειών στην λήψη κάποιων φαρμάκων (HapMap 2003).

Για να μπορέσουμε να βρούμε τους γενετικούς παράγοντες που παίζουν ρόλο στην εμφάνιση των παραπάνω στοιχείων, θα πρέπει να είμαστε σε θέση να αποκτήσουμε στα χέρια μας την πλήρη γενετική ακολουθία κάποιων ατόμων. Θα χρειαστούμε κάποια άτομα που να έχουν την ασθένεια, αλλά και κάποια υγιή. Κατόπιν θα πρέπει να συγκρίνουμε τις ακολουθίες αυτών που έχουν την ασθένεια, και αυτών που είναι υγιείς, για διαφορές στο γονιδίωμά τους. Μία τέτοια μέθοδος είναι δύσκολο να επιτευχθεί διότι το κόστος του να αποκτήσουμε την πλήρη ακολουθία του γονιδιώματος ενός ατόμου είναι μεγάλο (HapMap 2003). Το HapMap

προτείνει μια πιο σύντομη μέθοδο η οποία περιγράφεται παρακάτω ως «το πρόβλημα της εύρεσης του απλότυπου» (Lin, Cutler et al. 2002).

Ένα βασικό πρόβλημα των απλότυπων, είναι ότι αν έχουμε τον γονότυπο ενός ατόμου, είναι δύσκολο να καταλήξουμε σε συμπέρασμα για το ποιος θα είναι ο απλότυπός του. Στους διπλοειδείς οργανισμούς, όπου ανήκει και ο άνθρωπος, υπάρχουν δυο αντίγραφα κάθε χρωμοσώματος και κατά συνέπεια και δυο περιοχές ενδιαφέροντος. Ο απλότυπος ουσιαστικά είναι τα στοιχεία της περιοχής ενδιαφέροντος του ενός μόνο αντίγραφου, ενώ ο γονότυπος περιλαμβάνει τα στοιχεία και των δυο χρωμοσωμάτων. Στις ασθένειες οι οποίες επηρεάζονται από παραπάνω από ένα γονίδιο, είναι συνήθως πιο χρήσιμη η πληροφορία που μας δίνει ο απλότυπος συγκριτικά με τον γονότυπο. Η πληροφορία από την οποία αποτελείται ο απλότυπος, όπως αναφέραμε παραπάνω, είναι είτε η πλήρης DNA ακολουθία της περιοχής ενδιαφέροντος, είτε το σύνολο των μονοσημειακών πολυμορφισμών σε εκείνη την περιοχή. Η δεύτερη εκδοχή είναι αυτή η οποία κυριαρχεί, αλλά και η πιο χρήσιμη, επειδή περιέχει ακριβώς την πληροφορία που πρέπει να αναλυθεί χωρίς περιττά τμήματα.

Το πρόβλημα αυτό της εύρεσης του απλότυπου μπορεί να αντιμετωπιστεί και ως υπολογιστικό πρόβλημα (Gusfield 2000), με την δημιουργία γονοτυπικών πινάκων, πινάκων δηλαδή που θα περιέχουν πληροφορία για περιοχές ενδιαφέροντος του γενετικού υλικού. Στις θέσεις των πινάκων αυτών θα υπάρχουν οι αριθμοί 0, 1 και 2. Για θέσεις του πίνακα όπου η αντίστοιχες περιοχές στα ομόλογα χρωμοσώματα είναι σε ομόζυγη κατάσταση θα αντιστοιχούν οι αριθμοί 0 ή 1 (κάθε SNP μπορεί να έχει δυο πιθανά αλληλόμορφα), ενώ για ετερόζυγη κατάσταση το 2.



Εικόνα 8: Το υπολογιστικό πρόβλημα της εύρεσης του απλότυπου

Για ένα σύνολο n πινάκων θα έχουμε ως αποτέλεσμα ένα σύνολο από n δυαδικούς πίνακες, όπου κάθε δυαδικός πίνακας αντιστοιχεί σε έναν από τους αρχικούς. Αυτός ο δυαδικός πίνακας θα πρέπει να έχει 0 ή 1 και στις δυο θέσεις όπου ο αρχικός πίνακας είχε 0 ή 1. Επίσης για τον πίνακα που είχε το 2, θα πρέπει ο δυαδικός να έχει 0 στην πρώτη θέση και 1 στην δεύτερη ή αντίστροφα. Αυτοί οι δυαδικοί πίνακες μπορούν να γραφούν με διάφορους τρόπους, και δηλώνουν έτσι τις πιθανές μορφές των απλότυπων που μπορούν να προκύψουν. Για ένα άτομο με h ετερόζυγα σημεία προκύπτουν 2^{h-1} ζευγάρια που θα μπορούσαν να είναι η λύση του προβλήματος της εύρεσης του απλότυπου (εικόνα 8). Ωστόσο χωρίς επιπλέον επεξεργασία των δυαδικών πινάκων δεν μπορούμε να βρούμε ποιές από τις πιθανές λύσεις είναι οι σωστές. Συνεπώς υπάρχει η ανάγκη χρήσης κάποιων γενετικών μοντέλων (Stephens and Donnelly 2003) για να μας καθοδηγήσουν προς την λύση.

1.4.1 Τα γενετικά μοντέλα

Η εύρεση ενός απλότυπου μπορεί να γίνει και πειραματικά σε βιολογικά εργαστήρια (Clark 1990), αλλά αυτή η μέθοδος είναι αρκετά χρονοβόρα, απαιτεί πολύ εργασία και κοστίζει πολλά χρήματα. Για αυτό το λόγο χρησιμοποιούνται εναλλακτικά διαφόρων ειδών στατιστικές μέθοδοι, οι οποίες μπορούν αν χρησιμοποιηθούν σωστά να είναι σύντομες και έγκυρες, χωρίς μεγάλες χρηματικές απαιτήσεις (Lin and Fann 2009). Αυτές οι στατιστικές μέθοδοι μπορούν να χωριστούν σε τρεις κύριες κατηγορίες, τον αλγόριθμο του Clark, τον αλγόριθμο EM (expectation-maximization) και τις προσεγγίσεις του Bayes (Niu 2004). Επιπλέον έχουν προταθεί και άλλοι αλγόριθμοι πέρα από τους κλασικούς (Gusfield 2000).

Ο αλγόριθμος του Clark (Clark 1990), αρχίζει με μια λίστα γνωστών απλότυπων από τα ευκρινή άτομα (άτομα δηλαδή που οι γονότυποι τους δεν έχουν ή έχουν μόνο μια ετερόζυγη περιοχή). Στη συνέχεια η μέθοδος επιχειρεί να αναλύσει και τους γονότυπους από τα εναπομείναντα άτομα χρησιμοποιώντας τους απλότυπους που βρήκε αρχικά και το συμπλήρωμά τους, δηλαδή με χρήση του εναλλακτικού τους αλληλόμορφου σε κάθε θέση, καθώς και τις συχνότητες τους. Κάποιες φορές ο αλγόριθμος αποτυγχάνει να αναγνωρίσει τους απλότυπους για όλα τα άτομα, λόγω του ότι αν βρεθεί σε λάθος μορφή κάποιος απλότυπος, αυτό θα οδηγήσει σε σφάλμα στα επόμενα βήματα. Παρ όλα αυτά ο αλγόριθμος του Clark είναι η πρώτη στατιστική μέθοδος για την απόκτηση πληροφορίας απλότυπων και έτσι θεμελιωδώς σημαντικός. Σαν είσοδο ο αλγόριθμος δέχεται n γονοτυπικούς πίνακες μήκους m. Όπως είπαμε παραπάνω αυτοί οι πίνακες περιέχουν τους αριθμούς 0, 1 ή 2. Οι θέσεις των πινάκων που έχουν τους αριθμούς 0 ή 1 θεωρούνται «αποφασισμένες» ενώ αυτές που έχουν 2 θεωρούνται «διφορούμενες». Ένας πίνακας που δεν έχει «διφορούμενες» θέσεις ονομάζεται και αυτός «αποφασισμένος», ενώ ένας πίνακας με «διφορούμενες» θέσεις ονομάζεται «διφορούμενος». Για δυο

Εικόνα 9: Το λογότυπο της εφαρμογής Hap

διαφορετικούς πίνακες Α και Β η συγχώνευση τους δημιουργεί τον πίνακα ΑΒ ο οποίος χαρακτηρίζεται «διφορούμενος» και οι θέσεις όπου και ο Α και ο Β είχαν 0 ή 1, γίνονται 0 ή 1 αντίστοιχα. Οι θέσεις του ΑΒ όπου οι Α και Β είχαν (1 και 0) ή (0 και 1) αντίστοιχα παίρνουν την τιμή 2.

Η μέθοδος του Clark αρχίζει με την ταυτοποίηση κάθε πίνακα με καμία ή μια «διφορούμενη» θέση, μιας και στην περίπτωση που δεν υπάρχει καμία «διφορούμενη» θέση ο απλότυπος που προκύπτει είναι ο ίδιος ο γονοτυπικός πίνακας απαράλλακτος. Στην περίπτωση που έχουμε μια «διφορούμενη» θέση τότε προκύπτουν δυο απλοτυπικοί πίνακες. Οι απλότυποι που προκύπτουν ονομάζονται «αρχικά αποφασισμένοι πίνακες».

Το κύριο κομμάτι του αλγορίθμου του Clark επεκτείνεται με βάση τους «αρχικά αποφασισμένους πίνακες» και αποφασίζει για τους απομένοντες «διφορούμενους» πίνακες. Ο Clark πρότεινε τον ακόλουθο κανόνα που συνάγει τον επόμενο «αποφασισμένο» πίνακα, από έναν γονοτυπικό πίνακα Α και από έναν ήδη «αποφασισμένο» πίνακα Β. Ο «αποφασισμένος» πίνακας Β μπορεί να είναι είτε από τα δεδομένα εισόδου του αλγορίθμου, είτε από την προηγούμενη εφαρμογή αυτού του κανόνα.

Ο κανόνας αυτός ονομάζεται «κανόνας απόφασης» και είναι ο εξής: Αν υποθέσουμε ότι ο πίνακας Α είναι «διφορούμενος» με h διφορούμενες θέσεις , και R ένας «αποφασισμένος» πίνακας ο οποίος αντιστοιχεί σε μια από τις 2^h πιθανές αναλύσεις του πίνακα Α, τότε ο Α είναι η συγχώνευση του ενός αντιγράφου του «αποφασισμένου» πίνακα R και ενός άλλου «αποφασισμένου» πίνακα NR (ο οποίος καθορίζεται με συγκεκριμένο τρόπο). Όλες οι αποφασισμένες θέσεις του Α αντιγράφονται στον πίνακα NR αντίθετα από ότι σε σχέση με τον R. Έτσι όταν ο πίνακας NR χαρακτηριστεί «αποφασισμένος», προστίθεται στο πλήθος των αποφασισμένων πινάκων, και ο πίνακας Α διαγράφεται από το σύνολο των πινάκων.

Για παράδειγμα αν ο Α είναι 0212 και ο R είναι 0110, τότε ο NR είναι 0011. Η ερμηνεία αυτού είναι ότι αν οι δυο απλότυποι σε ένα άτομο είναι 0110 και 0011 τότε ο γονότυπος που παρατηρείται θα είναι 0212. Ο κανόνας απόφασης βγάζει συμπέρασμα για τον πίνακα 0212 με το σκεπτικό ότι ο 0110 είναι ένας απλότυπος στον πληθυσμό, για να αποφασίσει ότι ο 0011 είναι επίσης ένας απλότυπος του πληθυσμού. Όταν ο κανόνας απόφασης μπορεί να εφαρμοστεί για την εξαγωγή συμπεράσματος στον πίνακα NR από τους πίνακες Α και R, τότε μπορούμε να πούμε ότι ο R μπορεί να εφαρμοστεί στον Α. Είναι εύκολο να αποφανθούμε αν ένας

«αποφασισμένος» πίνακας R μπορεί να εφαρμοστεί σε έναν «διφορούμενο» πίνακα A: Ο R μπορεί να εφαρμοστεί στον A αν και μόνο αν ο A δεν περιέχει καμία θέση s, τέτοια ώστε οι αποφασισμένες τιμές του A και του R να διαφέρουν σε αυτή τη θέση s. Μια «αποφασισμένη» θέση s στον A της οποίας η τιμή διαφέρει από αυτήν στον R λέγεται ότι σταματάει την δυνατότητα εφαρμογής του R στον A. Για παράδειγμα ο 0110 δεν μπορεί να εφαρμοστεί στον 2012, γιατί η δεύτερη θέση εμποδίζει την εφαρμογή. Η διαδικασία αυτή από τον (Gusfield 2001), μας δίνει μια ιδέα για το πώς λειτουργεί ένας αλγόριθμος εύρεσης απλότυπων.

Ο αλγόριθμος αυτός για να βγάλει συμπέρασμα για ένα σύνολο γονότυπων, πρέπει πρώτα να αναγνωρίσει το αρχικά «αποφασισμένο» σύνολο, και στη συνέχεια εφαρμόζει τον κανόνα απόφασης επαναληπτικά, μέχρι να χαρακτηριστούν όλοι οι γονότυποι «αποφασισμένοι» ή να μην υπάρχουν άλλοι γονότυποι που να μπορούν να χαρακτηριστούν «αποφασισμένοι».

Ο αλγόριθμος ΕΜ αντί να καταλήξει σε έναν απλότυπο για κάθε άτομο, στοχεύει στην εκτίμηση της συχνότητας των απλότυπων με την μέγιστη πιθανοφάνεια. Αντιμετωπίζει τους γονότυπους σαν καταμετρημένα, αλλά ημιτελή δεδομένα με άγνωστους απλότυπους. Αρχίζοντας από μια αυθαίρετη συχνότητα απλότυπων, ενημερώνει αυτή την συχνότητα περιοδικά μέσα από κάθε ΕΜ κύκλο μέχρι να επιτευχθεί σύγκλιση.

Οι προσεγγίσεις του Bayes περιλαμβάνουν μια ομάδα από πολύπλοκες διαδικασίες που βασίζονται στις Monte Carlo μεθόδους (Sun, Greenwood et al. 2007). Αυτές οι μέθοδοι είναι ένα είδος υπολογιστικών αλγόριθμων, οι οποίοι βασίζονται στην επανάληψη τυχαίας δειγματοληψίας για να υπολογίσουν αποτελέσματα. Χρησιμοποιούνται συνήθως στην προσομοίωση φυσικών, μαθηματικών, αλλά και βιολογικών συστημάτων.

Την υλοποίηση όλων αυτών των μεθόδων καλείται να πραγματοποιήσει ο ηλεκτρονικός υπολογιστής, που ως γνωστόν, μπορεί να εκτελέσει μαθηματικές πράξεις με πολύ μεγάλη ταχύτητα. Έχουν δημιουργηθεί πολλές εφαρμογές οι οποίες βρίσκουν τον απλότυπο χρησιμοποιώντας τις παραπάνω μεθόδους που περιγράψαμε (Xu, Wu et al. 2004).

Εικόνα 10: Η εφαρμογή Haplo-i χρησιμοποιεί ένα μοντέλο του Bayes για την επίλυση του προβλήματος εύρεσης απλότυπων

Τα πιο γνωστά προγράμματα τα οποία χρησιμοποιούν μεθόδους του Bayes (Sun, Greenwood et al. 2007) είναι τα: HAPLOTYPER, HAP(εικόνα 9), PHASE και HAPLOi (εικόνα 10). Και τα τρία αυτά προγράμματα δίνουν σαν αποτέλεσμα απλότυπους για κάθε άτομο. Οι διαφορές τους βρίσκονται κυρίως στις υποθέσεις για την προηγούμενη πιθανοτική κατανομή και στην υλοποίηση του υπολογιστικού αλγορίθμου. Για παράδειγμα το PHASE υποθέτει μια προηγούμενη κατανομή που προήλθε από ένα προσεγγιστικό μίγμα, ενώ το HAP και το HAPLOTYPER υποθέτουν μια Dirichlet προηγούμενη κατανομή (Xing, Jordan et al. 2007) η οποία ανήκει σε μια οικογένεια πολυπαραγοντικών κατανομών.

Το PHASE βρέθηκε να έχει τα πιο ακριβή αποτελέσματα στην εύρεση του απλότυπου, τον υπολογισμό των συχνοτήτων και στις μελέτες συσχέτισης. Επιπλέον τα αποτελέσματά του δεν είχαν κανένα λάθος συγκριτικά με τους απλότυπους που προήλθαν από βιολογικά πειράματα. Οι άλλες δυο μέθοδοι του HAP και HAPLOTYPER είχαν επίσης χαμηλό ποσοστό λαθών. Συμπερασματικά και οι τρεις μέθοδοι έδωσαν συνεπή με την εργαστηριακή προσέγγιση αποτελέσματα, σε λιγότερο χρόνο και κόστος (Xu, Wu et al. 2004).

ΚΕΦΑΛΑΙΟ 2 – ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Ορισμός του προβλήματος

Σκοπός αυτής της εργασίας είναι αρχικά να διερευνήσουμε την συμβολή των πολυμορφισμών στην συμπεριφορά κοινών παθήσεων (όπως διαβήτης, καρκίνος, παθήσεις της καρδιάς κλπ). Αν δηλαδή κάποια σύνολα απλότυπων ή ακόμη και μεμονωμένοι απλότυποι που βρίσκονται στο γονιδίωμα των ανθρώπων, παίζουν ή όχι σημαντικό ρόλο στην εμφάνιση κάποιας τέτοιας ασθένειας. Επίσης θέλουμε να προτείνουμε κάποιες απλές μεθόδους για την ανάλυση αυτού του τύπου βιολογικών δεδομένων. Τα δεδομένα των απλότυπων δίνονται από μελέτες που χωρίζονται σε δυο κύριες κατηγορίες. Μελέτες με ασθενείς και μάρτυρες, και μελέτες με μεταβλητή συνεχούς τιμής.

2.2 Η εφαρμογή Stata

Το Stata είναι ένα στατιστικό πακέτο γενικής φύσεως το οποίο δημιουργήθηκε το 1985 από την εταιρία StataCorp και είναι ευρέως διαδεδομένο τόσο σε εταιρίες, όσο και σε ακαδημαϊκά κέντρα σε όλο τον κόσμο. Οι περισσότεροι χρήστες το προτιμούν για έρευνα, ειδικά σε τομείς όπως της οικονομίας, της κοινωνιολογίας και της επιδημιολογίας.

Οι δυνατότητες της εφαρμογής αυτής περιλαμβάνουν:

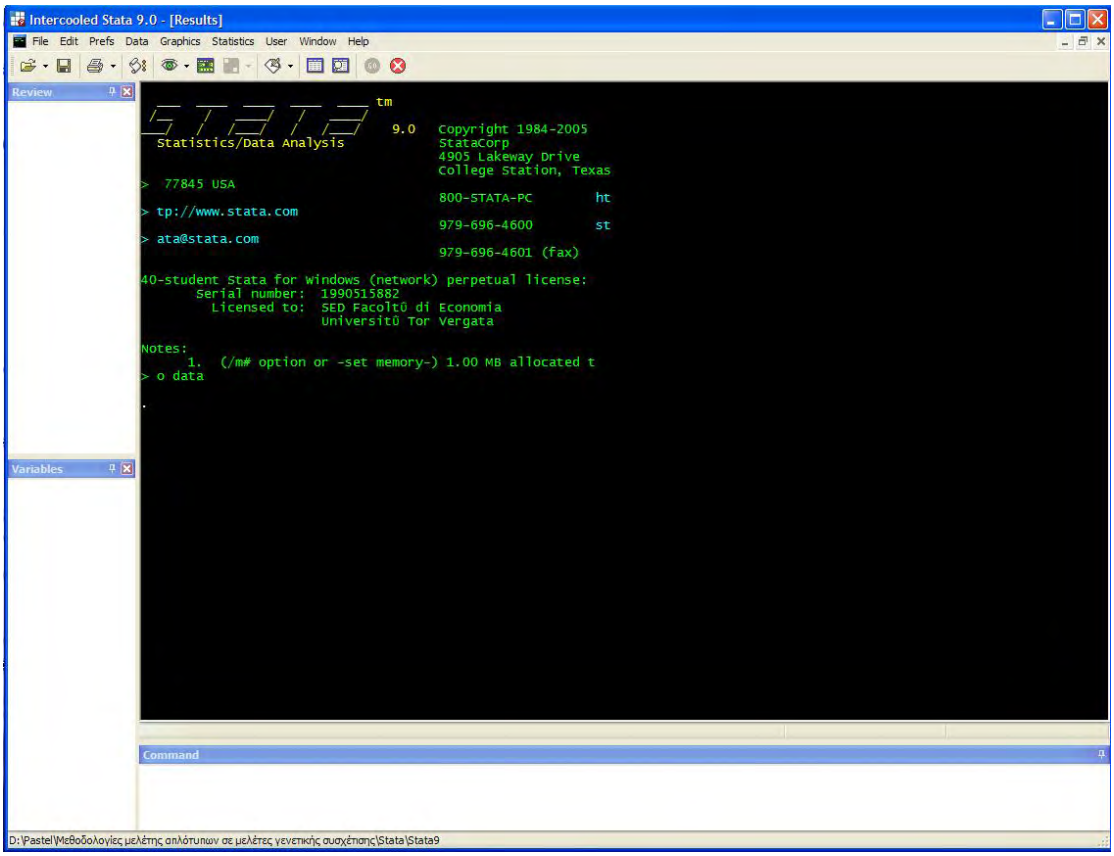
- Διαχείριση των δεδομένων
- Στατιστική ανάλυση
- Γραφικές απεικονίσεις
- Προσομοιώσεις
- Προγραμματισμός από τον χρήστη

Το όνομα Stata είναι μια συγχώνευση δυο λέξεων, statistics που σημαίνει στατιστική, και data, που σημαίνει δεδομένα. Υπάρχουν συνολικά τέσσερις εκδοχές του προγράμματος. Η πρώτη είναι κατασκευασμένη για πολυ-πύρινους υπολογιστές, η δεύτερη είναι ειδικά σχεδιασμένη για εξαιρετικά μεγάλες βάσεις δεδομένων, η τρίτη είναι η βασική έκδοση(και η πιο διαδεδομένη πιθανότατα) και τέλος η τέταρτη η οποία είναι για μικρού μεγέθους δεδομένα, κυρίως για εκπαιδευτική χρήση.

Στην κύρια οθόνη του προγράμματος υπάρχουν τέσσερα παράθυρα όπως φαίνεται και στην εικόνα 11, στο μεγαλύτερο από αυτά, το κεντρικό παράθυρο, εμφανίζονται τα αποτελέσματα των διεργασιών που εκτελεί η εφαρμογή (Results). Κάτω από το κύριο παράθυρο, είναι το παράθυρο των εντολών (Command), εκεί ο χρήστης πληκτρολογεί τις προς εκτέλεση εντολές. Τα δυο υπολειπόμενα παράθυρα έχουν βοηθητικό και πληροφοριακό ρόλο. Στο πάνω αριστερά, το επονομαζόμενο Review, καταγράφονται οι εντολές που έχουν εκτελεστεί από τον χρήστη μέχρι πρότινος. Και τέλος στο κάτω αριστερά, το επονομαζόμενο Variables βλέπουμε μια λίστα με τα ονόματα των υπαρχόντων μεταβλητών.

Παρόλο που υπάρχει το παράθυρο των εντολών και οι χρήστες εισάγουν τα δεδομένα και πραγματοποιούν στατιστικούς ελέγχους με την μορφή εντολών, η εφαρμογή παρέχει και εναλλακτικό τρόπο, χρησιμοποιώντας το κλασικό μενού επιλογών στο πάνω μέρος της εικόνας 11(File Edit Prefs κ.λ.π.).

Θα πρέπει να σημειωθεί πως για την εισαγωγή δεδομένων από το excel θα πρέπει οι μεταβλητές να είναι μια λέξη, ή εφόσον είναι δύο η περισσότερες να μην χωρίζονται με κενό, αλλά με το σύμβολο της κάτω παύλας «_».



Εικόνα 11: Η κύρια οθόνη του Stata

2.3 Εκτίμηση του Linkage disequilibrium

Στην πληθυσμιακή γενετική, ο όρος linkage disequilibrium που από εδώ και πέρα θα αναφερόμαστε ως LD, περιγράφει μη τυχαία συσχέτιση αλληλόμορφων σε δύο ή περισσότερα σημεία, τα οποία δεν βρίσκονται απαραίτητα στο ίδιο χρωμόσωμα. Δεν είναι το ίδιο με την γενετική συσχέτιση, η οποία περιγράφει την σχέση δύο ή περισσότερων θέσεων σε ένα χρωμόσωμα. Δύο περιοχές που βρίσκονται κοντά πάνω στην αλυσίδα του DNA είναι γενετικά συσχετισμένες με την έννοια ότι έχουν πολλές πιθανότητες μετά την διαδικασία του ανασυνδυασμού και της μείωσης να παραμείνουν και να κληρονομηθούν μαζί, στο ίδιο χρωμόσωμα.

Έτσι ο όρος LD περιγράφει μια κατάσταση στην οποία κάποιοι συνδυασμοί αλληλόμορφων ή γενετικών δεικτών παρατηρούνται λιγότερο ή περισσότερο σε έναν πληθυσμό από ότι θα ήταν αναμενόμενο από μια τυχαία διάταξη απλότυπων κάποιων αλληλόμορφων με βάση τις συχνότητές τους. Η μη-τυχαία συσχέτιση μεταξύ πολυμορφισμών σε διαφορετικές θέσεις μετριέται από τον βαθμό του LD. Αριθμητικά είναι η διαφορά μεταξύ των παρατηρηθέντων και των αναμενόμενων (υποθέτουμε τυχαία κατανομή) συχνοτήτων των αλληλόμορφων (Devlin and Risch 1995).

Ένα παράδειγμα είναι η εξάπλωση δύο σπάνιων ασθενειών στην Φιλανδία, όπου εκεί συγκριτικά με την υπόλοιπη Ευρώπη, η κυστική ίνωση εξαπλώνεται με αργό ρυθμό. Αντίθετα η έμφυτη χλωριούχος διάρροια (congenital chloride diarrhea) εξαπλώνεται γρήγορα. Και οι δύο ασθένειες οφείλονται σε μεταλλάξεις στο χρωμόσωμα 7, σε γειτονικά γονίδια. Κάτι τέτοιο μπορεί να σημαίνει διαφορές μεταξύ γενετικού υλικού Φιλανδών και υπόλοιπων Ευρωπαίων.

Ο βαθμός του LD επηρεάζεται από διάφορους παράγοντες όπως η γενετική συσχέτιση, ο ρυθμός του ανασυνδυασμού, ο ρυθμός των μεταλλάξεων και τη δομή του πληθυσμού. Για παράδειγμα, κάποιοι οργανισμοί όπως τα βακτήρια, τα οποία αναπαράγονται χωρίς μείωση καθότι απλοειδή, παρουσιάζουν LD επειδή απουσιάζει ο ανασυνδυασμός και συνεπώς οι περιοχές που συνδέονται παραμένουν σταθερές.

2.3.1 Είδη και μεθοδολογίες υπολογισμού του LD

Υπάρχουν διάφοροι τρόποι υπολογισμού της συσχέτισης όσον αφορά το LD. Για να κάνουμε τους υπολογισμούς, πρέπει πρώτα να ορίσουμε έναν πίνακα 2X2. Σύμφωνα με τους Devlin and Risch υποθέτουμε ότι έχουμε δύο περιοχές

ενδιαφέροντος στο γενετικό υλικό, και κάθε περιοχή έχει δύο αλληλόμορφα. Ένα αλληλόμορφο για την ασθένεια, και ένα κανονικό αλληλόμορφο τα οποία απομονώνονται στην πρώτη περιοχή ενδιαφέροντος. Τα άλλα δύο αλληλόμορφα τα οποία χρησιμοποιούνται ως γενετικοί δείκτες, απομονώνονται στην δεύτερη περιοχή (Devlin and Risch 1995). Σε αυτήν την περίπτωση τα δεδομένα μας μπορούν να πάρουν την μορφή του πίνακα 1. Παρατηρήστε τα σύνολα που καταγράφονται δεξιά και κάτω από τις απλοτυπικές συχνότητες, καθώς παίζουν σημαντικό ρόλο στους παρακάτω υπολογισμούς. Το n_{11} είναι ο αριθμός των απλότυπων στο δείγμα που φέρουν το αλληλόμορφο της ασθένειας για το A1. Το n_{A1} είναι ο συνολικός αριθμός των απλότυπων που φέρουν το A1 αλληλόμορφο, n_{α} είναι ο αριθμός των απλότυπων που φέρουν το αλληλόμορφο της ασθένειας και n είναι ο συνολικός αριθμός των απλότυπων του δείγματος.

Δείκτης	Αλληλόμορφο για την ασθένεια	Κανονικό αλληλόμορφο	
A1	n_{11}	n_{12}	n_{A1}
A2	n_{21}	n_{22}	n_{A2}
	n_{α}	n_{κ}	n

Πίνακας 1: Διάταξη συχνοτήτων των απλότυπων σε 2X2 πίνακα

Διαιρώντας αυτές τις ποσότητες με το n παίρνουμε τις αναμενόμενες συχνότητες και τις πιθανότητες p του δείγματος, (πίνακας 2). Αυτό μας βοηθάει σε διάφορους υπολογισμούς όπως για παράδειγμα, να βρούμε την πιθανότητα του να υπάρχει το αλληλόμορφο A1 στον απλότυπο, δεδομένου ότι έχουμε αλληλόμορφο της ασθένειας. Αυτό σημαίνει ότι $p_{A1/ασθένεια} = p_{11} / p_{\alpha}$. Ομοίως η πιθανότητα να έχουμε κανονικό αλληλόμορφο στον απλότυπο δεδομένου ότι υπάρχει μόνο το αλληλόμορφο A2 είναι $p_{κανονικό/A2} = p_{22} / p_{A2}$.

Βέβαια αυτές οι ποσότητες (p) είναι μόνο μια εκτίμηση κάποιων άγνωστων παραμέτρων, οι οποίες θα συμβολιστούν με π. Χρησιμοποιούμε τα π στους ακόλουθους ορισμούς, με την προϋπόθεση ότι αυτές οι άγνωστες ποσότητες υπολογίζονται από τις παρατηρούμενες ποσότητες του δείγματος.

Δείκτης	Αλληλόμορφο για την ασθένεια	Κανονικό αλληλόμορφο	
A1	p_{11}	p_{12}	p_{A1}
A2	p_{21}	p_{22}	p_{A2}
	p_{α}	p_{κ}	1

Πίνακας 2: Κάθε πιθανότητα προκύπτει από τα στοιχεία του πίνακα 1 διαιρούμενα με το n

Ο κοινός παρονομαστής για κάθε μέτρο του LD είναι, όπως αναφέραμε και πριν, η διαφορά μεταξύ παρατηρηθέντων και αναμενόμενων (υπό ανεξαρτησία) αριθμών των απλότυπων που έχουν το αλληλόμορφο της ασθένειας και βρίσκονται στον δείκτη A1, ή τις ισοδύναμες εκφράσεις της σχέσης 1.

$$\begin{aligned} D &= \pi_{11} - \pi_{\mathrm{A1}}\pi_{\alpha} = \pi_{22} - \pi_{\mathrm{A2}}\pi_{\kappa} \\ &= -\pi_{12} + \pi_{\mathrm{A1}}\pi_{\kappa} = -\pi_{21} + \pi_{\mathrm{A2}}\pi_{\alpha} \\ &= \pi_{11}\pi_{22} - \pi_{12}\pi_{21} \end{aligned}$$

(1)

Σύμφωνα με τους Hill and Weir, το πιο ευρέως χρησιμοποιούμενο μέτρο για τον υπολογισμό του LD είναι το τετράγωνο του μέτρου Δ.

$$\Delta = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\left(\pi_{\mathrm{A1}}\pi_{\mathrm{A2}}\pi_{\alpha}\pi_{\kappa}\right)^{1/2}}$$

(2)

ή αλλιώς το Δ^2 (Hill and Weir 1994).

Άλλο ένα μέτρο που χρησιμοποιείται συχνά είναι το D΄.

$$D' = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{A1}\pi_{\kappa}, \pi_{\alpha}\pi_{A2})} D > 0$$

$$D' = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{A1}\pi_{\alpha}, \pi_{\kappa}\pi_{A2})} D < 0$$

(3)

όπου η ποσότητα στον παρονομαστή είναι το μέγιστο D που θα μπορούσε να επιτευχθεί με τα δεδομένα του πίνακα. Αυτά τα δυο μέτρα (D', Δ) σχετίζονται με στατιστικές μεθόδους υπολογισμού της συσχέτισης. Συγκεκριμένα το Δ είναι ο συντελεστής συσχέτισης για έναν πίνακα 2X2. Επίσης το Δ είναι ανάλογο του r (σχέση 4) για έναν πίνακα 2X2.

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{(\hat{m}_{ij}(1 - p_{Ai})(1 - p_{\alpha/\kappa}))^{1/2}}$$

(4)

όπου το $p_{\alpha/\kappa}$ υπολογίζεται ανάλογα με το αν οι δείκτες i, j αντιστοιχούν σε στήλη με το αλληλόμορφο της ασθένειας ή με το κανονικό αλληλόμορφο, και το m_{ij} είναι ο αναμενόμενος αριθμός στο κελί ij .

Άλλο ένα μέτρο συσχέτισης που χρησιμοποιείται στην επιδημιολογία για τον υπολογισμό του LD είναι το δ^*, το οποίο ορίζεται ως:

$$\delta^* = \frac{\pi_{A1} + (\varphi - 1)}{1 + \pi_{A1}(\varphi - 1)}$$

(5)

όπου $\varphi = \{\pi_{11} / \pi_{A1}\} / \{\pi_{21} / \pi_{A2}\}$ είναι ο σχετικός κίνδυνος ή αλλιώς odds ratio.

Το δ^* μπορούμε να το βρούμε μετά από κάποιες αλλαγές και στην μορφή της σχέσης 6.

$$\delta = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{22}}$$

(6)

η οποία είναι μια προσέγγιση του κινδύνου του πληθυσμού σε μια μεγαλύτερη κλίμακα από αυτή των πληθυσμιακών συχνοτήτων τους.

Αυτό το μέτρο δ^* δεν είναι καινούργιο καθώς, όταν η ασθένεια είναι σπάνια και οι απλότυποι έχουν δειγματοληφθεί τυχαία, τότε ισχύει ότι $\delta^* = D'$:

$$\delta^* = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{\mathrm{A}2}}$$

(7)

όμως

$$D' = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\min(\pi_{\mathrm{A}1}\pi_{\kappa}, \pi_{\alpha}\pi_{\mathrm{A}2})}$$
$$= \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{\mathrm{A}2}}$$

(8)

βλέπουμε ότι τα δύο κλάσματα των σχέσεων (7) και (8) συμπίπτουν.

Άλλο ένα μέτρο το οποίο ενδείκνυται για χρήση σε μελέτες ασθενών-μαρτύρων είναι το d, το οποίο ορίζεται με την παρακάτω σχέση:

$$d = \frac{\pi_{11}}{\pi_{\alpha}} - \frac{\pi_{12}}{\pi_{\kappa}} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{\kappa}}$$

(9)

Επίσης υπάρχουν και άλλα φυσικά μέτρα για το LD, όπως είναι ο σχετικός λόγος των πιθανοτήτων (odds ratio) αλλά και το Q του Yule. (Clegg, Kidwell et al. 1976; Nei and Li 1980; Olson and Wijsman 1994).

Ο λόγος των πιθανοτήτων ορίζεται ως:

$$OR = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{10}$$

με εύρος τιμών $[0, +\infty]$, ενώ το Q ως:

$$Q = \frac{OR - 1}{OR + 1} = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}} \tag{11}$$

με εύρος τιμών $[-1, +1]$. Το τελευταίο κλάσμα της εξίσωσης 11 μας δείχνει την σχέση του Q με το δ. Στην πραγματικότητα οι αριθμητές των Δ, D', δ, d και Q είναι όλοι ίσοι με το D, καθώς αυτά τα μέτρα διαφέρουν μόνο κατά τους παρονομαστές τους, οι οποίοι έχουν την χρήση της τυποποίησης του D. Στον πίνακα 3 βλέπουμε συνοπτικά τα προαναφερθέντα μέτρα.

Σύμβολο	Τύπος
Δ	$\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{(\pi_{A1}\pi_{A2}\pi_{\alpha}\pi_{\kappa})^{1/2}}$
D΄	$\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\min(\pi_{A1}\pi_{\kappa},\pi_{\alpha}\pi_{A2})}$
δ	$\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{22}}$
d	$\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\pi_{\alpha}\pi_{\kappa}}$
Q	$\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}+\pi_{12}\pi_{21}}$

Πίνακας 3: Μια σύνοψη των μέτρων υπολογισμού του LD

2.4 Μελέτες με ασθενείς και μάρτυρες

Οι μελέτες με ασθενείς και μάρτυρες είναι ένα είδος επιδημιολογικών μελετών. Χρησιμοποιούνται για την αναγνώριση κάποιων παραγόντων που επηρεάζουν ασθένειες, είτε θετικά είτε αρνητικά. Η μελέτη γίνεται πάνω σε έναν πληθυσμό ατόμων τα οποία αποτελούνται από άτομα που έχουν κάποια ασθένεια και από άτομα υγιή, χωρίς όμως να παρουσιάζουν σημαντικές διαφορές από τους ασθενείς (Fallin, Cohen et al. 2001), έχουν δηλαδή τις ίδιες συνθήκες ζωής και έχουν επιλεγεί με τα ίδια κριτήρια.

Αυτού του είδους οι μελέτες είναι ένα πολύ φθηνό εργαλείο που χρησιμοποιείται ευρέως και με μεγάλη ευκολία, από μικρές ομάδες ή ερευνητές. Το βασικό πλεονέκτημα είναι ότι δεν απαιτείται ιδιαίτερος εργαστηριακός εξοπλισμός καθώς βασίζονται στην στατιστική και όχι στην εργαστηριακή ανάλυση. Έχουν οδηγήσει σε σημαντικές ανακαλύψεις, αλλά έχουν το αρνητικό χαρακτηριστικό ότι μπορεί να υπάρχουν παράγοντες που δεν ξέρουμε, οι οποίοι επηρεάζουν κατά ένα ποσοστό το αποτέλεσμα.

Παρακάτω ακολουθεί μια εικόνα ενός πίνακα ασθενών-μαρτύρων (εικόνα 12). Στις στήλες του πίνακα διαχωρίζονται οι ασθενείς και οι μάρτυρες, ενώ στις γραμμές βλέπουμε τους διάφορους παράγοντες που μπορεί να σχετίζονται με την ασθένεια. Κάποιοι βασικοί παράγοντες είναι το κάπνισμα, ο τόπος κατοικίας, η ηλικία και η ύπαρξη άλλων μορφών καρκίνου. Κάτω από την γραμμή κάθε παράγοντα, παρατηρούμε και το p-value, το οποίο δείχνει αν απορρίπτουμε ή δεχόμαστε την μηδενική υπόθεση, η οποία είναι ότι δεν υπάρχει συσχέτιση παράγοντα-ασθένειας.

Table 1. Selected Characteristics of Lung Cancer Cases and Controls With Interview Data Available, the EAGLE Study, Lombardy, Italy, 2002–2005[a,b]

	Women				Men			
	Cases		Controls		Cases		Controls	
	No.	%	No.	%	No.	%	No.	%
Total participants enrolled	448		500		1,652		1,620	
Interviewed	406	100.0	499	100.0	1,537	100.0	1,617	100.0
Area of residence								
Milan	288	70.9	349	69.9	987	64.2	1,089	67.3
Monza	24	5.9	23	4.6	109	7.1	94	5.8
Brescia	47	11.6	53	10.6	203	13.2	194	12.0
Pavia	21	5.2	37	7.4	107	7.0	92	5.7
Varese	26	6.4	37	7.4	131	8.5	148	9.2
	$P = 0.55$				$P = 0.17$			
Age, years (mean (SD))	64.8 (10.1)		64.1 (10.1)		66.8 (7.9)		65.8 (8.1)	
	$P = 0.32$				$P < 0.001$			
Educational level								
None	21	5.2	24	4.8	91	5.9	66	4.1
Elementary school	128	31.5	143	28.7	625	40.7	431	26.7
Middle school	134	33.0	158	31.7	424	27.6	455	28.1
High school	104	25.6	135	27.1	314	20.4	441	27.3
University	19	4.7	39	7.8	83	5.4	224	13.9
	$P = 0.35$				$P < 0.001$			
No. of jobs held								
1	166	40.9	168	33.7	375	24.4	370	22.9
2	96	23.7	158	31.7	404	26.3	356	22.0
3	77	19.0	82	16.4	305	19.8	356	22.0
4	30	7.4	49	9.8	194	12.6	226	14.0
≥5	37	9.1	42	8.4	259	16.9	309	19.1
	$P = 0.03$				$P = 0.02$			

	Women				Men			
	Cases		Controls		Cases		Controls	
	No.	%	No.	%	No.	%	No.	%
Cigarette smoking								
Never	103	25.4	282	56.5	29	1.9	397	24.6
Former (quit >6 months ago)	116	28.6	110	22.0	723	47.0	799	49.4
Current	187	46.1	107	21.4	785	51.1	420	26.0
Unknown	0	0.0	0	0.0	0	0.0	1	0.1
	$P < 0.001$				$P < 0.001$			
Cigarette pack-years (mean (SD))	24.3 (23.1)		7.2 (13.5)		50.9 (28.7)		22.1 (23.2)	
	$P < 0.001$				$P < 0.001$			
Other cancer(s)[c]								
No	336	82.8	448	89.8	1,306	85.0	1,473	91.1
Yes	70	17.2	51	10.2	231	15.0	144	8.9
	$P = 0.002$				$P < 0.001$			
Lung cancer morphology								
Adenocarcinoma	220	54.2			582	37.9		
Squamous cell carcinoma	45	11.1			459	29.9		
Large cell carcinoma	28	6.9			61	4.0		
Non-small-cell carcinoma NOS	34	8.4			142	9.2		
Small cell carcinoma	38	9.4			157	10.2		
Other	26	6.4			65	4.2		
Not available	15	3.7			71	4.6		
	$P < 0.001$							

Abbreviations: EAGLE, Environment And Genetics in Lung cancer Etiology; NOS, not otherwise specified; SD, standard deviation.

[a] *P* values were derived from the chi-square test (categorical variables) or Student's *t* test (continuous variables).

[b] Percentages may not add to 100.0 because of rounding.

[c] Primary cancer(s) (previously or newly diagnosed) other than lung cancer.

Εικόνα 12: Χαρακτηριστικά ασθενών-μαρτύρων από μια μελέτη για την σχέση μεταξύ καρκίνου του πνεύμονα και επαγγέλματος (Consonni, De Matteis et al. 2010)

Σε αυτήν εδώ την εργασία τα δεδομένα έχουν την ίδια μορφή με την παραπάνω μελέτη. Η μόνη διαφορά είναι ότι οι παράγοντες κινδύνου δεν είναι ένα σύνολο από διατροφικές συνήθειες, ηλικίες κ.λ.π., αλλά από ένα σύνολο διαφορετικών απλότυπων. Έτσι η έρευνα επικεντρώνεται στο να διαπιστώσουμε κάποιες διαφορές ανάμεσα σε άτομα με διαφορετικό απλότυπο ώστε να μπορέσουμε να κάνουμε μια εκτίμηση για τον βαθμό που επηρεάζεται η ασθένεια από αυτόν. Συνεπώς για έναν πίνακα ασθενών και μαρτύρων θα έχουμε στην πρώτη στήλη τους ασθενείς (cases) και στην δεύτερη στήλη τους μάρτυρες (controls). Ενώ στην πρώτη γραμμή, τον απλότυπο τύπου 1, στην δεύτερη τον απλότυπο τύπου 2 κ.ο.κ.

	ασθενείς (cases)	μάρτυρες (controls)
απλότυπος 1	252	248
απλότυπος 2	289	234
…		
…		
…		
…		
…		
…		
απλότυπος n	234	654

Πίνακας 4: Η κύρια μορφή των δεδομένων στις μελέτες με ασθενείς και μάρτυρες

Με βάση τον πίνακα 4 μπορούμε να πραγματοποιήσουμε διάφορες στατιστικές διεργασίες με το πρόγραμμα Stata, οι οποίες θα μας βοηθήσουν να εξάγουμε κάποια συμπεράσματα. Τα δεδομένα σε αυτήν την εργασία έχουν την παρακάτω μορφή (εικόνα 13):

Microsoft Excel - testing

Αρχείο Επεξεργασία Προβολή Εισαγωγή Μορφή Εργαλεία Δεδομένα Παράθυρο Βοήθεια

Arial 10

A8 haplotype

	A	B	C	D	E	F	G	H	I	J	K
8	haplotype	haplotype_nr	frequency	frequency_percent	case	cases_sum	x1	x2	x3	x4	x5
9	GGGCA	1	1147	0,621	1	1827	0	0	0	0	0
10	GGTTG	2	298	0,1613	1	1827	0	0	1	1	1
11	GGGCG	3	163	0,0881	1	1827	0	0	0	0	1
12	AGGCA	4	114	0,0675	1	1827	1	0	0	0	0
13	GAGCA	5	105	0,0568	1	1827	0	1	0	0	0
14	GGGCA	1	529	0,6347	0	0	0	0	0	0	0
15	GGTTG	2	132	0,1589	0	0	0	0	1	1	1
16	GGGCG	3	73	0,088	0	0	0	0	0	0	1
17	AGGCA	4	51	0,0615	0	0	1	0	0	0	0
18	GAGCA	5	44	0,0526	0	0	0	1	0	0	0
19											

Εικόνα 13: Η μορφή των δεδομένων που εισάγονται στο Stata

Αφού αντιγράψουμε τα δεδομένα από τον πίνακα του excel, για να τα εισάγουμε στο Stata πληκτρολογούμε την εντολή «edit» στην κύρια οθόνη του, και στην συνέχεια στο παράθυρο που εμφανίζεται κάνουμε επικόλληση.

2.4.1 Ο έλεγχος χ^2 του Pearson για την ανεξαρτησία δυο δειγμάτων

Υπάρχουν δύο έλεγχοι του Pearson. Ο πρώτος αναφέρεται σαν τεστ καλής προσαρμογής, και ο δεύτερος (ο οποίος χρησιμοποιείται και αναλύεται εδώ) αναφέρεται ως τεστ ανεξαρτησίας. Με αυτό τον έλεγχο, έχοντας δυο σύνολα δεδομένων, μπορούμε να ελέγξουμε κατά πόσο αυτά τα σύνολα είναι στατιστικώς συνδεδεμένα. Σαν μηδενική υπόθεση έχουμε το γεγονός ότι τα δυο σύνολα είναι στατιστικώς ανεξάρτητα. Τα δεδομένα οργανώνονται σε έναν πίνακα με r γραμμές και c στήλες, έπειτα υπολογίζεται για κάθε κελί του πίνακα η ακόλουθη θεωρητική συχνότητα:

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N} \tag{12}$$

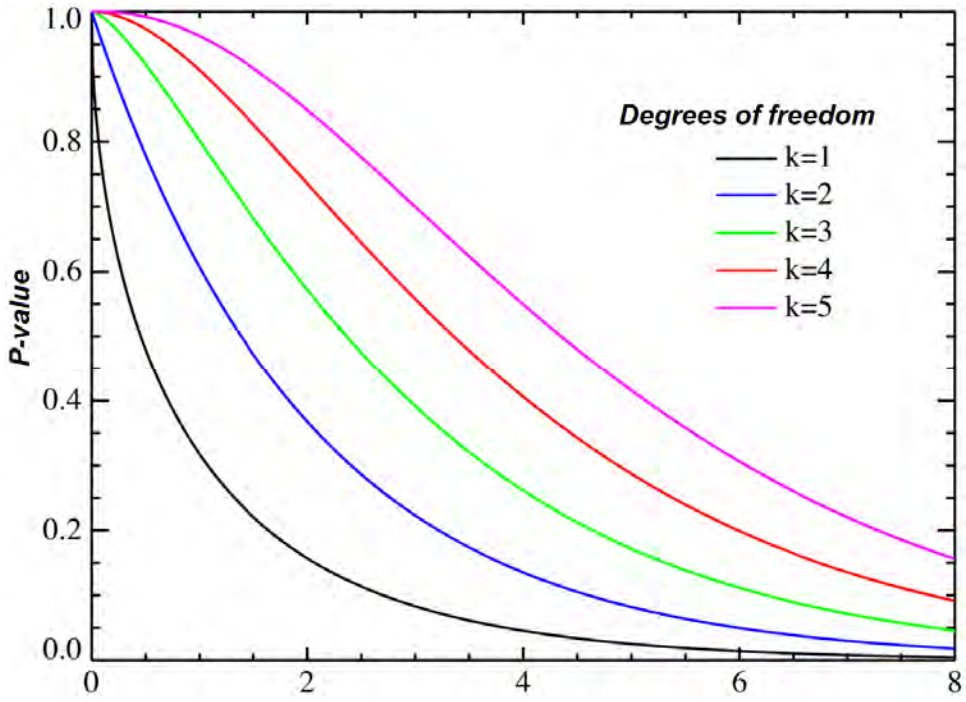
Αυτή η προσαρμογή μειώνει τους βαθμούς ελευθερίας κατά $p = r + c - 1$.

Τέλος υπολογίζουμε την τιμή του X^2 χρησιμοποιώντας την σχέση:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{13}$$

Ο αριθμός των βαθμών ελευθερίας είναι ίσος με τον αριθμό των κελιών του πίνακα, μείον τον αριθμό p.

Αν από αυτό το τεστ έχουμε σαν αποτέλεσμα πιθανότητα X^2 ίση ή μικρότερη από 0.05, τότε η μηδενική υπόθεση απορρίπτεται και μπορούμε να πούμε ότι υπάρχει τυχαία συσχέτιση μεταξύ των δυο συνόλων. Αυτήν την πιθανότητα την υπολογίζουμε από την γραφική παράσταση της κατανομής X^2 (εικόνα 14).



Εικόνα 14: Η κατανομή X^2

Παραδείγματος χάρη για να πραγματοποιήσουμε αυτόν τον έλεγχο σε δεδομένα όπως αυτά της εικόνας 13, αφού πρώτα τα εισάγουμε στο Stata, εκτελούμε την εντολή:

```
tabulate haplotype_nr case [fweight=frequency], row col chi2
```

Η εντολή `tabulate` δημιουργεί πίνακα συχνοτήτων μεταξύ των μεταβλητών `haplotype_nr` και `case`. Επιπλέον ορίζουμε ότι θέτουμε ως βάρος την συχνότητα, με την παράμετρο `[fweight=frequency],` και ζητάμε να γίνει ο έλεγχος X^2 με αναφορά ταυτόχρονα των σχετικών συχνοτήτων για κάθε γραμμή και κάθε στήλη (`row col`).

Τα αποτελέσματα της οποίας είναι τα ακόλουθα (εικόνα 15):

Key			
	frequency		
	row percentage		
	column percentage		

haplotype_nr	case 0	case 1	Total
1	3,283	2,412	5,695
	57.65	42.35	100.00
	44.08	48.70	45.92
2	2,623	1,626	4,249
	61.73	38.27	100.00
	35.22	32.83	34.26
3	674	453	1,127
	59.80	40.20	100.00
	9.05	9.15	9.09
4	504	273	777
	64.86	35.14	100.00
	6.77	5.51	6.27
5	161	51	212
	75.94	24.06	100.00
	2.16	1.03	1.71
6	48	33	81
	59.26	40.74	100.00
	0.64	0.67	0.65
7	155	105	260
	59.62	40.38	100.00
	2.08	2.12	2.10
Total	7,448	4,953	12,401
	60.06	39.94	100.00
	100.00	100.00	100.00

Pearson chi2(6) = 48.6233 Pr = 0.000

Εικόνα 15: Αποτελέσματα του χ^2 του Pearson

Στα αποτελέσματα αυτά παρατηρούμε ότι ο χ^2 έχει τιμή περίπου 48.6 και το p-value είναι κάτω από 0.05, αυτό δείχνει ότι υπάρχει τυχαία συσχέτιση μεταξύ του συνόλου των απλότυπων, και της ασθένειας. Ο έλεγχος αυτός έχει εισαγωγικό χαρακτήρα, καθώς όταν βλέπουμε πως υπάρχει τυχαία συσχέτιση μεταξύ των στοιχείων, μας προτρέπει να ασχοληθούμε παραπάνω με την συγκεκριμένη μελέτη. Στην αντίθετη περίπτωση κάτι τέτοιο θα ήταν πιθανότατα περιττό.

2.4.2 Εκτίμηση μέγιστης πιθανοφάνειας

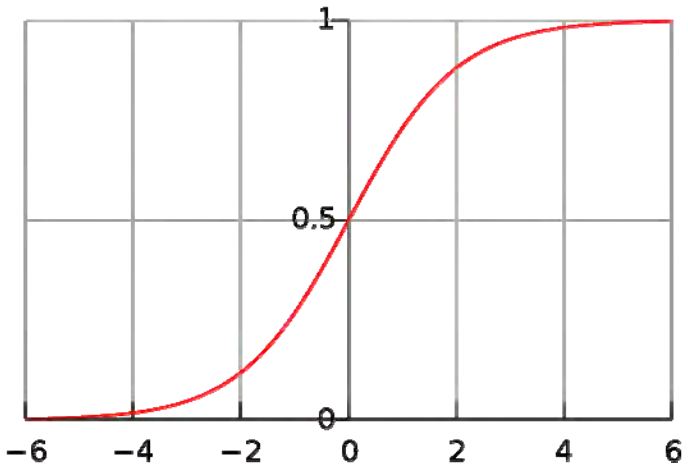
Η εκτίμηση μέγιστης πιθανοφάνειας είναι μια μέθοδος που χρησιμοποιούμε για την προσαρμογή ενός στατιστικού μοντέλου σε κάποια δεδομένα, καθώς επίσης και να κάνουμε μια εκτίμηση για τις παραμέτρους του μοντέλου αυτού. Για παράδειγμα ας υποθέσουμε ότι θέλουμε να ασχοληθούμε με το ύψος ενός πληθυσμού. Έχουμε ένα δείγμα ατόμων, αλλά όχι ολόκληρο των πληθυσμό (της χώρας, της κοινότητας κλπ), και καταγράφουμε τα ύψη τους. Θα πρέπει να υποθέσουμε ότι τα ύψη ακολουθούν κανονική κατανομή με κάποια άγνωστη μέση τιμή και διακύμανση. Τότε ο μέσος όρος του δείγματος είναι η εκτιμήτρια μέγιστης

πιθανοφάνειας του μέσου όρου ολόκληρου του πληθυσμού, όπως και η διακύμανση του δείγματος είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της διακύμανσης του πληθυσμού.

2.4.3 Λογιστική παλινδρόμηση

Στη στατιστική η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πραγματοποίησης ενός γεγονότος, με την προσαρμογή κάποιων δεδομένων σε μια λογιστική συνάρτηση. Είναι ένα γενικευμένο γραμμικό μοντέλο που χρησιμοποιείται για την εξαγωγή αποτελεσμάτων διωνυμικού τύπου. Δηλαδή ισχύει-δεν ισχύει. Για την πρόβλεψη αυτή γίνεται χρήση διαφόρων μεταβλητών, είτε κατηγορικών, είτε αριθμητικών (Agresti 2002).

Για παράδειγμα η πιθανότητα για κάποιον να πάθει έμφραγμα μέσα σε μια συγκεκριμένη περίοδο, μπορεί να προβλεφθεί από τις γνώσεις που έχουμε για την ηλικία, το φύλο και τον δείκτη μάζας σώματος. Αυτό το μοντέλο χρησιμοποιείται σε μεγάλο βαθμό στις ιατρικές και κοινωνικές επιστήμες, αλλά και σε εφαρμογές του marketing όπως την πρόβλεψη των τάσεων του πελάτη για την αγορά ενός προϊόντος. Για να εξηγήσουμε την λογιστική παλινδρόμηση, πρέπει πρώτα να εξηγήσουμε την λογιστική συνάρτηση (εικόνα 16).



Εικόνα 16: Η συνάρτηση f(z) με διάφορες τιμές του z στον άξονα των x

Σαν είσοδο της συνάρτησης έχουμε τον αριθμό z, και σαν έξοδο έχουμε το f(z). Ο z παίρνει τιμές από το $-\infty$ έως το $+\infty$ ενώ το πεδίο τιμών της f(z) είναι το (0,1). Η μεταβλητή z απεικονίζει την έκθεση σε ένα σύνολο από ανεξάρτητες

μεταβλητές, ενώ η f(z) απεικονίζει την πιθανότητα ενός συγκεκριμένου αποτελέσματος, με βάσει τις παραπάνω μεταβλητές.

Ο τύπος της συνάρτησης είναι ο εξής:

$$f(z) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{e^{z}+1}$$

(14)

Επιπλέον το z εκφράζει την συνολική συνεισφορά όλων των ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν στο μοντέλο, και είναι γνωστό και ως logit.

$$z = \text{logit}(p)$$

(15)

Το logit ενός αριθμού p από 0 μέχρι 1 δίνεται από τον παρακάτω τύπο:

$$\log it(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

(16)

Η μεταβλητή z ορίζεται συνήθως σύμφωνα με τον παρακάτω τύπο:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots + \beta_k x_k$$

(17)

όπου το στοιχείο β_0 ονομάζεται στοιχείο αναφοράς και τα β_1, β_2, β_3 ..... β_k ονομάζονται συντελεστές παλινδρόμησης των x_1, x_2, x_3 ..... x_k αντίστοιχα.

Το στοιχείο αναφοράς είναι η τιμή του z όταν η τιμή όλων των ανεξάρτητων μεταβλητών είναι μηδέν (για παράδειγμα η τιμή του z όταν απεικονίζει την πιθανότητα καρδιακής προσβολής για ένα άτομο, χωρίς επιβαρυντικούς παράγοντες).

Κάθε ένας από τους συντελεστές παλινδρόμησης παριστάνει το μέγεθος της συνεισφοράς του αντίστοιχου επιβαρυντικού παράγοντα.

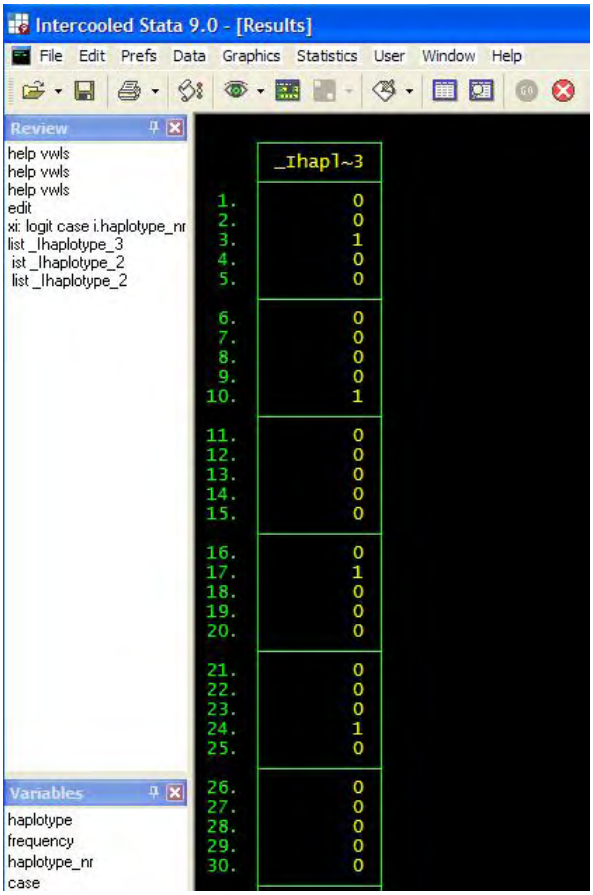
Ένας θετικός συντελεστής παλινδρόμησης δείχνει ότι ο αντίστοιχος παράγοντας αυξάνει την πιθανότητα πραγματοποίησης του αποτελέσματος, ενώ ένας αρνητικός συντελεστής παλινδρόμησης σημαίνει ότι ο αντίστοιχος παράγοντας μειώνει την πιθανότητα πραγματοποίησης του αποτελέσματος.

Επιπλέον ένας μεγάλος συντελεστής παλινδρόμησης σημαίνει ότι ο αντίστοιχος παράγοντας επηρεάζει κατά πολύ την πιθανότητα πραγματοποίησης του αποτελέσματος, ενώ ένας συντελεστής παλινδρόμησης κοντά στο μηδέν έχει μικρή επιρροή στην πιθανότητα πραγματοποίησης του αποτελέσματος.

Για να προσαρμόσουμε αυτό το μοντέλο στα δεδομένα μας, πρέπει να σκεφτούμε πως έχουμε μια κατηγορική μεταβλητή (case) για τους ασθενείς/μάρτυρες, η οποία παίζει το ρόλο της εξαρτημένης μεταβλητής, και μετά έχουμε τους απλότυπους που επηρεάζουν την κατάσταση της μεταβλητής αυτής θετικά ή αρνητικά, συνεπώς χρησιμοποιούνται ως ανεξάρτητες μεταβλητές. Έτσι μπορούμε να δημιουργήσουμε μια πιθανότητα εμφάνισης της ασθένειας, λαμβάνοντας υπόψη την επιρροή των απλότυπων. Έχουμε λοιπόν την πιθανότητα $\pi_j = P(y_j = 1 | j)$ του να είναι κάποιος ασθενής (δηλαδή y=1), έχοντας τον j απλότυπο. Το καλύτερο σενάριο θα ήταν να διαλέξουμε τον πιο κοινό απλότυπο (δηλαδή τον απλότυπο με την μεγαλύτερη συχνότητα) ως το στοιχείο αναφοράς, και να φτιάξουμε μια μεταβλητή για κάθε έναν από τους υπόλοιπους. Η μεταβλητή αυτή θα παίρνει τιμές $z_j = 1$ για τον j απλότυπο, και 0 σε όλες τις άλλες περιπτώσεις. Στην εικόνα 17 βλέπουμε ένα παράδειγμα τέτοιας μεταβλητής για τον απλότυπο 3, τον τρίτο δηλαδή κατά σειρά απλότυπο. Παρατηρούμε ότι μόνο η τρίτη θέση έχει τον αριθμό 1. Αν αναφερόμασταν πχ στον απλότυπο 6, το 1 θα ήταν στην έκτη κατά σειρά θέση. Εδώ οι απλότυποι είναι 7 στο σύνολο και απλά η διαδικασία επαναλαμβάνεται πολλές φορές. Αυτό το μοντέλο λοιπόν μπορεί να πραγματοποιηθεί ως εξής, (Wallenstein, Hodge et al. 1998):

$$\log it(\pi_j) = \log it\left[P(y_j = 1 | j) \right] = \beta_0 + \sum_{j=2}^{r} \beta_j z_j$$

(18)



Εικόνα 17: Η μορφή της μεταβλητής που κωδικοποιεί τον απλότυπο 3

Για να εκτελέσουμε αυτόν τον έλεγχο στο Stata, αφού εισάγουμε τα δεδομένα, πρέπει να πληκτρολογήσουμε την παρακάτω εντολή:

```
xi : logit case i.haplotype_nr [fweight=frequency]
```

Με το `xi` στην αρχή της εντολής δηλώνουμε ότι θέλουμε να χρησιμοποιήσουμε μια κατηγορική μεταβλητή η οποία θα έχει έναν δείκτη. Αυτή θα είναι η `haplotype_nr` όπου βάζοντας μπροστά της το `i.` δηλώνουμε στο Stata ότι αυτή θα είναι η μεταβλητή μας. Πέρα από αυτά με την εντολή `logit` δηλώνουμε πως θέλουμε να κάνουμε λογιστική παλινδρόμηση με εξαρτημένη μεταβλητή την `case` και ανεξάρτητη την `i.haplotype_nr`. Επιπλέον θέλουμε να συμπεριλάβουμε τις συχνότητες των δεδομένων με την παράμετρο `[fweight=frequency]`.
Και τα αποτελέσματα που εμφανίζονται στην οθόνη είναι τα εξής:

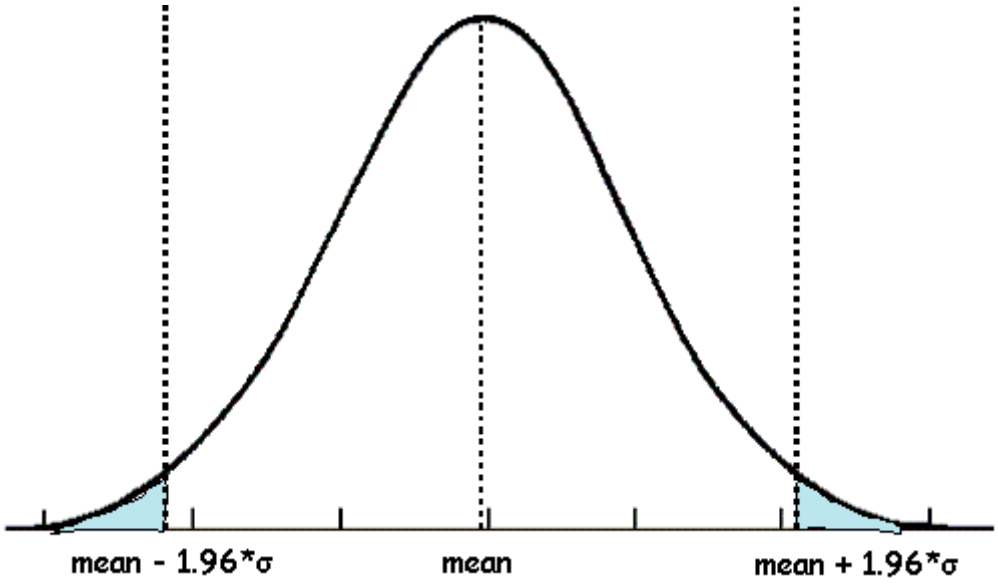
```
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)

Iteration 0:   log likelihood =  -8343.008
Iteration 1:   log likelihood = -8318.0226
Iteration 2:   log likelihood = -8317.8958
Iteration 3:   log likelihood = -8317.8958

Logistic regression                                     Number of obs   =      12401
                                                        LR chi2(6)      =      50.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -8317.8958                             Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_Ihaplotyp~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_Ihaplotyp~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_Ihaplotyp~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_Ihaplotyp~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_Ihaplotyp~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
------------------------------------------------------------------------------
```

Εικόνα 18: Αποτελέσματα της λογιστικής παλινδρόμησης

Στην εικόνα 18 βλέπουμε έξι στήλες. Η πρώτη στήλη με το όνομα case δείχνει τις διάφορες ανεξάρτητες μεταβλητές οι οποίες επηρεάζουν το συνολικό αποτέλεσμα της παλινδρόμησης. Παρατηρούμε ότι μια μεταβλητή δεν υπάρχει, πράγμα αναμενόμενο καθώς μια από τις μεταβλητές επιλέγεται ως το «στοιχείο αναφοράς» (_Ihaplotype_1). Στην δεύτερη στήλη έχουμε τους συντελεστές παλινδρόμησης που αντιστοιχούν στις ανεξάρτητες μεταβλητές.

Στην επόμενη στήλη βλέπουμε έναν υπολογισμό της τυπικής απόκλισης. Στη συνέχεια βλέπουμε τις τιμές της ποσότητας z, η οποία όπως αναφέραμε παραπάνω είναι μια λογαριθμική διαφορά (σχέσεις 15, 16) και ισούται με το στοιχείο logit(p). Οι μεταβολές αυτής της ποσότητας δείχνουν ουσιαστικά την πορεία των υπολογισμών του μοντέλου καθώς υπολογίζεται η συνεισφορά κάθε συντελεστή παλινδρόμησης.

Στη συνέχεια βλέπουμε τις τιμές του p-value. Ανάλογα με το διάστημα εμπιστοσύνης, αν η τιμή του p-value είναι μικρότερη από έναν οριακό αριθμό, τότε πρέπει να απορρίψουμε την αρχική υπόθεση που κάναμε στην αρχή του στατιστικού ελέγχου. Σε αυτήν εδώ την εργασία χρησιμοποιούμε διάστημα εμπιστοσύνης 95% πράγμα που σημαίνει πως ο οριακός αριθμός είναι το 0.05. Η περιοχή μεταξύ των διακεκομμένων γραμμών καλύπτει το 95% της κατανομής, συνεπώς υπάρχει 95% πιθανότητα να έχουμε μια τιμή εντός αυτού του πλαισίου (εικόνα 19). Τέλος έχουμε την στήλη του διαστήματος εμπιστοσύνης. Αποτελείται από δυο αριθμούς, το κάτω όριο και το πάνω όριο.



Εικόνα 19: Το διάστημα εμπιστοσύνης και τα όριά του

Υπάρχουν και άλλοι τρόποι απεικόνισης των αποτελεσμάτων. Όπως ξέρουμε, η παλινδρόμηση χρησιμοποιεί κατά την διάρκεια των υπολογισμών τον λόγο των πιθανοτήτων (σχέση 10). Για κάθε συντελεστή παλινδρόμησης ισχύει ότι $\exp(\beta_i) = OR$.

Για παράδειγμα στην συγκεκριμένη μελέτη, ο απλότυπος 1 έχει επιλεχθεί από το Stata ως στοιχείο αναφοράς. Αυτό σημαίνει πως οι επόμενοι απλότυποι θα εξεταστούν ως προς την απόκλισή τους από το στοιχείο αναφοράς, πράγμα που γίνεται με χρήση του πίνακα 5.

	case	control
haplotype 2	1626	2623
haplotype 1	2412	3283

Πίνακας 5: Απλότυποι σε σχέση με ασθένεια

Υπολογίζοντας τον λόγο των πιθανοτήτων του πίνακα προκύπτει ότι:

OR = 0.843754

Όταν ο λόγος είναι ίσος με την μονάδα, σημαίνει πως τα άτομα που έχουν τον απλότυπο 1, έχουν τον ίδιο κίνδυνο να εμφανίσουν την ασθένεια με τα άτομα που έχουν τον απλότυπο 2. Όταν ο λόγος είναι μεγαλύτερος από την μονάδα, τότε τα άτομα που έχουν τον απλότυπο 2 έχουν μεγαλύτερη πιθανότητα να εμφανίσουν την ασθένεια σε σύγκριση με αυτούς που έχουν τον απλότυπο 1. Τέλος όταν ο λόγος είναι

μικρότερος από την μονάδα, τότε τα άτομα που έχουν τον απλότυπο 2 έχουν μικρότερη πιθανότητα να εμφανίσουν την ασθένεια σε σύγκριση με αυτούς που έχουν τον απλότυπο 1.

Η ίδια διαδικασία θα επαναληφθεί για όλους τους απλότυπους της μελέτης και συνεπώς προκύπτουν και οι αντίστοιχοι λόγοι πιθανοτήτων.

Αξίζει να σημειώσουμε πως κάθε φορά συγκρίνεται κάθε απλότυπος με τον πρώτο, ο οποίος έχει και την μεγαλύτερη συχνότητα, και συνεπώς έχουμε μια εκτίμηση του πόσο αλλάζουν οι πιθανότητες ασθένειας για διαφορετικούς απλότυπους σε σύγκριση με τον πιο κοινό.

Με την ακόλουθη εντολή, το Stata μπορεί να μας εμφανίσει τους σχετικούς λόγους πιθανοτήτων (Odds ratio) που έχει υπολογίσει, κάτι που μπορεί να βοηθήσει στην καλύτερη ερμηνεία των αποτελεσμάτων:

```
xi : logit case i.haplotype_nr [fweight=frequency], or
```

Δεν αλλάζει κάτι στην λειτουργία της λογιστικής παλινδρόμησης, απλά προσθέτουμε σαν επιπλέον επιλογή το `or`.

Το αποτέλεσμα φαίνεται στην εικόνα 20.

```
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)

Iteration 0:   log likelihood =  -8343.008
Iteration 1:   log likelihood = -8318.0226
Iteration 2:   log likelihood = -8317.8958
Iteration 3:   log likelihood = -8317.8958

Logistic regression                               Number of obs   =      12401
                                                  LR chi2(6)      =      50.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -8317.8958                       Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |    .843754   .0349465    -4.10   0.000     .7779664    .9151047
_Ihaplotyp~3 |   .9148121   .0607533    -1.34   0.180     .8031617    1.041983
_Ihaplotyp~4 |   .7372685   .0588261    -3.82   0.000     .6305347    .8620698
_Ihaplotyp~5 |   .4311594   .0702382    -5.16   0.000     .3133087    .5933396
_Ihaplotyp~6 |   .9357639   .2130904    -0.29   0.771     .5988696    1.462178
_Ihaplotyp~7 |    .922043    .119135    -0.63   0.530     .7157633    1.187772
------------------------------------------------------------------------------
```

Εικόνα 20: Αποτελέσματα της λογιστικής παλινδρόμησης με Odds Ratio

Παρατηρούμε ότι η στήλη με τους συντελεστές παλινδρόμησης έχει αντικατασταθεί από τους λόγους πιθανοτήτων.

Μετά την εκτέλεση της παλινδρόμησης, μπορούμε να εκτελέσουμε έναν έλεγχο πάνω στους συντελεστές (coefficients) για να δούμε ποιος από αυτούς

διαφέρει από τους υπόλοιπους, και να έχουμε μια ποσοτική εκτίμηση για την διαφορά αυτή. Ο έλεγχος αυτός έχει τις εξής υποθέσεις:

$H_0 =$ Όλοι οι συντελεστές είναι ίσοι με 0

$H_1 =$ Κάποιοι ή όλοι οι συντελεστές δεν είναι 0

Για να εκτελέσουμε αυτή τη διαδικασία στο Stata πρέπει να πληκτρολογήσουμε την παρακάτω εντολή:

```
testparm _Ihaplotype_*
```

Έτσι δηλώνουμε ότι θέλουμε να κάνουμε τον έλεγχο αυτό στους συντελεστές της μεταβλητής `_Ihaplotype_*` η οποία προέκυψε από την παλινδρόμηση. Το σύμβολο '*' δηλώνει πως θέλουμε όλα τα περιεχόμενα της μεταβλητής αυτής.
και τα αποτελέσματα φαίνονται στην εικόνα 21.

```
( 1)  _Ihaplotype_2 = 0
( 2)  _Ihaplotype_3 = 0
( 3)  _Ihaplotype_4 = 0
( 4)  _Ihaplotype_5 = 0
( 5)  _Ihaplotype_6 = 0
( 6)  _Ihaplotype_7 = 0

           chi2(  6) =   47.57
         Prob > chi2 =    0.0000
```

Εικόνα 21: Τα αποτελέσματα του testparm

Ουσιαστικά συγκρίνουμε όλους τους συντελεστές με τον ίδιο αριθμό (εδώ το 0) και μετράμε την απόστασή τους από αυτόν. Στην περίπτωση αποδοχής της H_0 οι διαφορές των συντελεστών δεν είναι μεγάλες και μπορούμε να συμπεράνουμε ότι δεν έχουμε κάποιον απλότυπο που να επηρεάζει κατά πολύ την ασθένεια. Αντίθετα στην H_1 βλέπουμε ότι υπάρχουν διαφορές, και αυτό μας προτρέπει να ερευνήσουμε ποιος ή ποιοι απλότυποι είναι που διαφέρουν. Επιπλέον υπολογίζεται και ο X^2 του Pearson, που δηλώνει το πόσο ισχυρή είναι η διαφορά και η ισότητα των συντελεστών αντίστοιχα.

Αυτός ο έλεγχος μπορεί να γίνει για διάφορους συνδυασμούς απλότυπων και δεν είναι υποχρεωτικό να έχουμε μια εκτίμηση μόνο για το σύνολό τους.

Παραδείγματος χάρη μπορούμε να ελέγξουμε ένα ζευγάρι απλότυπων, μια τριάδα κλπ. Σε αυτήν την περίπτωση η σύνταξη της εντολής αλλάζει και γίνεται έτσι:

```
test _Ihaplotype_3 _Ihaplotype_7
```

Έτσι ελέγχουμε την ομοιότητα των συντελεστών του απλότυπου 3 και του 7, βάζοντας απλά τον αντίστοιχο αριθμό στην θέση του '*' που είδαμε πιο πριν.

Γίνεται σαφές ότι αυτός ο επιπλέον έλεγχος είναι απαραίτητος, διότι στα αποτελέσματα της εικόνας 21 βλέπουμε πως στο σύνολό τους οι απλότυποι διαφέρουν όλοι μεταξύ τους. Στην συνέχεια όμως θα δούμε πως ο 3 και ο 7 δεν έχουν στατιστικώς μεγάλη διαφορά (εικόνα 22) καθώς δεχόμαστε την μηδενική υπόθεση.

```
( 1)  _Ihaplotype_3 = 0
( 2)  _Ihaplotype_7 = 0

           chi2(  2) =    2.07
         Prob > chi2 =    0.3560
```

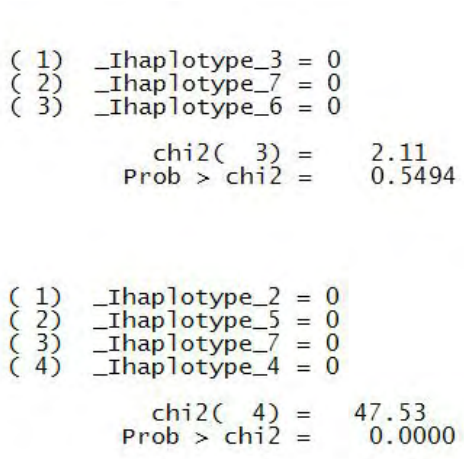
Εικόνα 22: Αποτελέσματα του test για δυο μεταβλητές

Βασικό κριτήριο πάντα για την αποδοχή ή απόρριψη της υπόθεσης είναι το p-value που όπως βλέπουμε στον πρώτο έλεγχο (εικόνα 21) είναι μικρότερο από 0.05 και συνεπώς δεν υπάρχει ισότητα, ενώ στον δεύτερο έλεγχο βλέπουμε τελικά ότι με p-value= 0.3560 οι απλότυποι 3 και 7 είναι στατιστικά ισοδύναμοι.

Φυσικά μπορούμε να κάνουμε αυτόν τον έλεγχο με όσες από τις μεταβλητές που έχουν αντίστοιχο συντελεστή παλινδρόμησης επιθυμούμε, πχ:

```
test _Ihaplotype_3 _Ihaplotype_7 _Ihaplotype_6
test _Ihaplotype_2 _Ihaplotype_5 _Ihaplotype_7 _Ihaplotype_4
```

Με αποτέλεσμα την εικόνα 23:



Εικόνα 23: αποτελέσματα του test για 3 και 4 μεταβλητές

2.4.4 Πολυωνυμική λογιστική παλινδρόμηση

Η πολυωνυμική λογιστική παλινδρόμηση είναι κατάλληλη για χρήση σε κατηγορικές, εξαρτημένες μεταβλητές, χρησιμοποιείται δηλαδή για την ανάλυση των σχέσεων μεταξύ μη – μετρικών εξαρτημένων μεταβλητών με μετρικές ή διχοτομικές ανεξάρτητων μεταβλητών. Συγκρίνει πολλές ομάδες μέσα από ένα συνδιασμό δυαδικών λογιστικών παλινδρομήσεων. Αυτό συμβαίνει επειδή όταν οι κατηγορίες είναι δυο, η διαδικασία είναι ίδια ακριβώς με την δυαδική λογιστική παλινδρόμηση που αναφέρουμε παραπάνω. Το πολυωνιμικό αυτό μοντέλο, είναι μια άμεση επέκταση των λογιστικών μοντέλων, και ο τρόπος υπολογισμού του έχει κάποια κοινά στοιχεία με αυτά (Chen and Kao 2006).

Ας υποθέσουμε ότι μια εξαρτημένη κατηγορική μεταβλητή έχει Μ κατηγορίες. Μια από τις τιμές της (συνήθως η πρώτη, η τελευταία, ή αυτή με την μεγαλύτερη συχνότητα) ορίζεται ως η κατηγορία αναφοράς. Αυτή η κατηγορία στην συνέχεια θα αποτελέσει κοινό παρονομαστή για όλες τις υπόλοιπες. Η πιθανότητα της σχέσης με άλλες κατηγορίες, συγκρίνεται με την πιθανότητα σχέσης με την κατηγορία αναφοράς.

Για μια εξαρτημένη μεταβλητή με Μ κατηγορίες, απαιτείται ο υπολογισμός Μ-1 εξισώσεων, μια για κάθε κατηγορία εκτός της κατηγορίας αναφοράς, για να περιγράψουμε την σχέση της εξαρτημένης μεταβλητής με τις ανεξάρτητες.

Έτσι, αν υποθέσουμε ότι για κατηγορία αναφοράς επιλέγεται η πρώτη, τότε για m = 2…..Μ, έχουμε:

$$\ln \frac{P(Y_i = m)}{P(Y_i = 1)} = a_m + \sum_{k=1}^{K} \beta_{mk} X_{ik} = Z_{mi}$$

(19)

Συνεπώς για κάθε περίπτωση θα έχουμε Μ-1 υπολογισμένες λογαριθμικές ποσότητες, μια για κάθε κατηγορία, σε σχέση με την κατηγορία αναφοράς.

Αξίζει να σημειωθεί πως όταν m=1, έχουμε σαν αποτέλεσμα στην σχέση 19 το ln(1) = 0 = Z_{11}.

Όταν έχουμε εξαρτημένη μεταβλητή με παραπάνω από δυο κατηγορίες, ο υπολογισμός των πιθανοτήτων γίνεται κάπως πιο πολύπλοκος από την λογιστική παλινδρόμηση. Για m = 2……M, έχουμε:

$$P(Y_i = m) = \frac{\exp(Z_{mi})}{1 + \sum_{m=2}^{M} \exp(Z_{mi})}$$

(20)

και για την κατηγορία αναφοράς:

$$P(Y_i = 1) = \frac{1}{1 + \sum_{m=2}^{M} \exp(Z_{mi})}$$

(21)

Συνοπτικά η διαδικασία έχεις ως εξής: παίρνουμε τους Μ-1 λογάριθμους που υπολογίσαμε, και τους υψώνουμε ως εκθέτες με βάση τον αριθμό e. Μόλις γίνει αυτό, ο υπολογισμός των πιθανοτήτων είναι εύκολη υπόθεση. Αυτή η μέθοδος βασίζεται στην πιθανότητα επιρροής από παράγοντες, δεδομένης της κατάστασης (ασθενής ή μάρτυρας), όπως εφαρμόζεται στις μελέτες ασθενών και μαρτύρων. Οι απλότυποι χρησιμοποιούνται ως εξαρτημένες μεταβλητές και η κατάσταση (ασθενής/μάρτυρας) λαμβάνεται ως κατηγορία αναφοράς (base outcome) σε αυτήν την μέθοδο (McCullagh 1999). Είναι εύκολο να δούμε ότι οι συντελεστές της παλινδρόμησης (β_j) είναι προσεγγίσεις του σχετικού λόγου των πιθανοτήτων καθώς $e^{\beta_j} = OR$.

Φυσικά ο πρώτος συντελεστής β_1=0 στην περίπτωση που χρησιμοποιείται ως κατηγορία αναφοράς. Αυτό το μοντέλο προτάθηκε για την ανάλυση δεδομένων απλότυπων πρώτα από τους (Chen and Kao 2006).

Για να εκτελέσουμε αυτόν τον έλεγχο στο Stata, αφού εισάγουμε τα δεδομένα, πρέπει να πληκτρολογήσουμε την παρακάτω εντολή:

```
xi: mlogit case i.haplotype_nr [fweight=frequency]
```

Με το `xi` στην αρχή της εντολής δηλώνουμε ότι θέλουμε να χρησιμοποιήσουμε μια κατηγορική μεταβλητή η οποία θα έχει έναν δείκτη. Αυτή θα είναι η `haplotype_nr` όπου βάζοντας μπροστά της το `i.` δηλώνουμε στο Stata ότι αυτή θα είναι η μεταβλητή μας. Πέρα από αυτά με την εντολή `mlogit` δηλώνουμε πως θέλουμε να κάνουμε πολυωνυμική λογιστική παλινδρόμηση με εξαρτημένη μεταβλητή την `case` και ανεξάρτητη την `i.haplotype_nr`. Επιπλέον θέλουμε να συμπεριλάβουμε τις συχνότητες των δεδομένων με την παράμετρο `[fweight=frequency]`.

Και τα αποτελέσματα εμφανίζονται στην εικόνα 24.

```
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)

Iteration 0:   log likelihood =  -8343.008
Iteration 1:   log likelihood = -8318.0226
Iteration 2:   log likelihood = -8317.8958
Iteration 3:   log likelihood = -8317.8958

Multinomial logistic regression                 Number of obs     =      12401
                                                LR chi2(6)        =      50.22
                                                Prob > chi2       =     0.0000
Log likelihood = -8317.8958                     Pseudo R2         =     0.0030

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_Ihaplotyp~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_Ihaplotyp~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_Ihaplotyp~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_Ihaplotyp~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_Ihaplotyp~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
------------------------------------------------------------------------------
(case==0 is the base outcome)

.
```

Εικόνα 24: Αποτελέσματα της πολυωνυμικής παλινδρόμησης

Παρατηρούμε ότι τα αποτελέσματα είναι ίδια με αυτά της προηγούμενης μεθόδου. Αυτό είναι αναμενόμενο καθώς υπάρχουν πολλές ομοιότητες.

Και σε αυτήν την περίπτωση μπορούμε να απεικονίσουμε τους σχετικούς λόγους των πιθανοτήτων με την παρακάτω εντολή:

```
xi: mlogit case i.haplotype_nr [fweight=frequency], rrr
```

Με το `rrr` (relative risk ratio) σαν πρόσθετη επιλογή, δηλαδή ο λόγος σχετικού κινδύνου, έχουμε το αποτέλεσμα της εικόνας 25.

```
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)

Iteration 0:   log likelihood =  -8343.008
Iteration 1:   log likelihood = -8318.0226
Iteration 2:   log likelihood = -8317.8958
Iteration 3:   log likelihood = -8317.8958

Multinomial logistic regression                 Number of obs   =      12401
                                                LR chi2(6)      =      50.22
                                                Prob > chi2     =     0.0000
Log likelihood = -8317.8958                     Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |    .843754   .0349465    -4.10   0.000     .7779664    .9151047
_Ihaplotyp~3 |   .9148121   .0607533    -1.34   0.180     .8031617    1.041983
_Ihaplotyp~4 |   .7372685   .0588261    -3.82   0.000     .6305347    .8620698
_Ihaplotyp~5 |   .4311594   .0702382    -5.16   0.000     .3133087    .5933396
_Ihaplotyp~6 |   .9357639   .2130904    -0.29   0.771     .5988696    1.462178
_Ihaplotyp~7 |    .922043    .119135    -0.63   0.530     .7157633    1.187772
------------------------------------------------------------------------------
(case==0 is the base outcome)
```

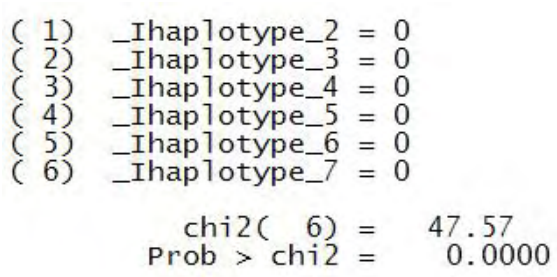
Εικόνα 25: Αποτελέσματα πολυωνυμικής παλινδρόμησης με relative risk ratio

Και πάλι βλέπουμε την στήλη με τους συντελεστές να γίνεται λόγος σχετικού κινδύνου για κάθε απλότυπο.

Στη συνέχεια μπορούμε να εκτελέσουμε τον έλεγχο ισότητας των συντελεστών με την παρακάτω εντολή:

```
test [1]
```

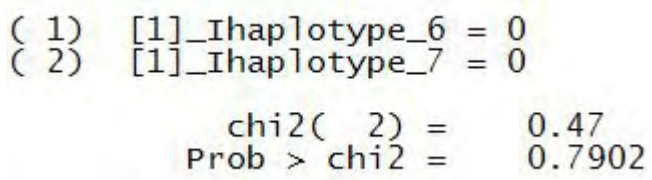
Και τα αποτελέσματα στην εικόνα 26 είναι ακριβώς ίδια με της λογιστικής παλινδρόμησης, απλώς αντί για την εντολή `testparm` χρησιμοποιούμε την `test [1]` και δεν είναι απαραίτητο να συμπεριλάβουμε την μεταβλητή με τους συντελεστές στην σύνταξη.



Εικόνα 26: Αποτελέσματα του test [1]

Επίσης μπορούμε όπως αναφερθήκαμε και παραπάνω να εκτελέσουμε αυτόν τον έλεγχο όχι μόνο για το σύνολο, αλλά και για κάποιες ομάδες απλότυπων. Η μόνη διαφορά φαίνεται στην σύνταξη της εντολής, όπου βάζουμε άνω και κάτω τελεία πριν τις μεταβλητές που επιθυμούμε, σε αντίθεση με την λογιστική παλινδρόμηση:

```
test [1]: _Ihaplotype_6 _Ihaplotype_7
```



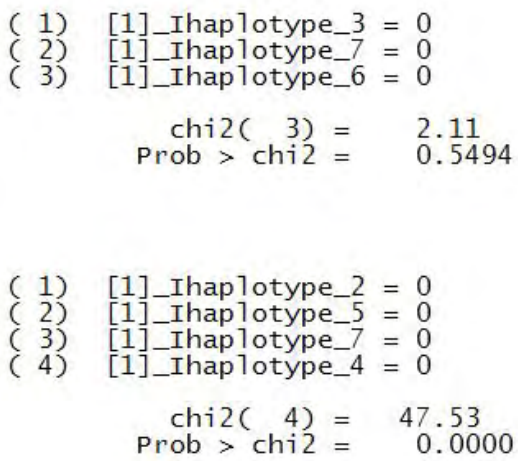
Εικόνα 27: test[1] για δυο μεταβλητές

Βλέπουμε ότι ο απλότυπος 7 είναι ίδιος με τον 6 (εικόνα 27), και μπορούμε να πούμε ότι οι απλότυποι 6 και 7 επηρεάζουν στην ίδια περίπου ένταση το αποτέλεσμα της παλινδρόμησης και κατ' επέκταση την ασθένεια. Όσο αυξάνεται ο αριθμός των ελέγχων που κάνουμε, τόσο καλύτερη εικόνα έχουμε για το τι συμβαίνει.

Φυσικά μπορούμε να κάνουμε αυτόν τον έλεγχο με όσες από τις μεταβλητές που έχουν αντίστοιχο συντελεστή παλινδρόμησης επιθυμούμε, πχ:

```
test [1]: _Ihaplotype_3 _Ihaplotype_7 _Ihaplotype_6
test [1]: _Ihaplotype_2 _Ihaplotype_5 _Ihaplotype_7 _Ihaplotype_4
```

Με τα ακόλουθα αποτελέσματα (εικόνα 28):

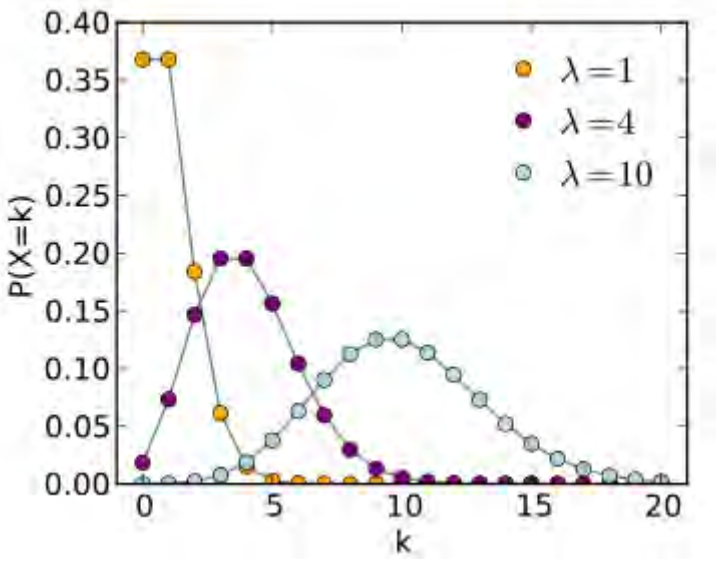


Εικόνα 28: Αποτελέσματα του test [1] για 3 και 4 μεταβλητές

2.4.5 Παλινδρόμηση Poisson

Αυτού του είδους η παλινδρόμηση σχετίζεται με την κατανομή Poisson (εικόνα 29), η οποία είναι μια κατανομή που εκφράζει την πιθανότητα εμφάνισης ενός αριθμού γεγονότων σε συγκεκριμένο χρόνο. Τα γεγονότα αυτά είναι ανεξάρτητα το ένα με το άλλο. Αν ο αναμενόμενος αριθμός των γεγονότων είναι λ, τότε η πιθανότητα να έχουμε ακριβώς n γεγονότα $(n = 0,1,2,...)$ είναι:

$$f(n,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{22}$$



Εικόνα 29: Η συνάρτηση μάζας πιθανότητας της κατανομής Poisson

Η παλινδρόμηση Poisson είναι μια μορφή παλινδρόμησης που χρησιμοποιείται για να μοντελοποιήσει απαριθμήσιμα δεδομένα και πίνακες ενδεχομένων. Απαριθμήσιμα δεδομένα, είναι τα δεδομένα στα οποία οι παρατηρήσεις παίρνουν μη-μηδενικές ακέραιες τιμές, και προκύπτουν από μέτρημα, όχι κατηγοριοποίηση. Οι πίνακες ενδεχομένων χρησιμοποιούνται για να καταγράψουν και να αναλύσουν τη σχέση μεταξύ δυο ή περισσοτέρων μεταβλητών.

Το μοντέλο χρησιμοποιεί μια μεταβλητή Y η οποία ακολουθεί κατανομή Poisson, καθώς και τον λογάριθμο της αναμενόμενης τιμής της μεταβλητής Y. Τα παραπάνω στοιχεία μπορούν να μοντελοποιηθούν από έναν γραμμικό συνδυασμό από

άγνωστες παραμέτρους. Το μοντέλο αυτό ονομάζεται και λογαριθμικό-γραμμικό (log-linear) μοντέλο, ειδικά όταν χρησιμοποιείται με πίνακες ενδεχομένων.
Η πιο απλή περίπτωση με μια ανεξάρτητη μεταβλητή x, έχει την εξής μορφή:

$$\log(E(Y)) = a + bx$$

(23)

Αν y_i είναι ανεξάρτητες παρατηρήσεις με αντίστοιχες τιμές x_i για την μεταβλητή x τότε τα α και b μπορούν να υπολογιστούν από την μέγιστη πιθανοφάνεια αν ο αριθμός των ευδιάκριτων x είναι τουλάχιστον 2.

Η παλινδρόμηση Poisson είναι κατάλληλη για διαβαθμιζόμενα δεδομένα, όπου η διαβάθμιση είναι ένα μέτρημα γεγονότων που λαμβάνουν χώρα σε μια συγκεκριμένη μονάδα παρατηρήσεων, διαιρούμενη από μια μονάδα μέτρησης της «έκθεσης» αυτής (exposure). Για παράδειγμα οι βιολόγοι μετρούν τον αριθμό των ειδών των δέντρων σε ένα δάσος όπου η διαβάθμιση θα είναι ο αριθμός των ειδών ανά τετραγωνικό χιλιόμετρο ή αυτοί που ασχολούνται με την δημογραφική επιστήμη μπορούν να μοντελοποιούν ρυθμούς θανάτων ανά γεωγραφικές περιοχές, ως τον αριθμό των θανάτων διαιρούμενο από άτομα/χρόνο.

Γενικότερα οι ρυθμοί των γεγονότων μπορούν να υπολογιστούν ως γεγονότα ανά μονάδα χρόνου. Στα παραπάνω παραδείγματα η «έκθεση» είναι αντίστοιχα το τετραγωνικό χιλιόμετρο, τα άτομα/χρόνο και ο χρόνος. Στην παλινδρόμηση Poisson αυτό αντιμετωπίζεται ως ένα αντιστάθμισμα, όπου η μεταβλητή της «έκθεσης» εισάγεται στην δεξιά πλευρά της εξίσωσης και με μια παραμετροποίηση στην λογαριθμική ποσότητα (log(exposure)) περιορίζεται μέχρι το 1.

$$\log(E(Y)) = \log(\text{exposure}) + a + bx$$

(24)

το οποίο γίνεται:

$$\log(E(Y)) - \log(\text{exposure}) = \log(\frac{E(Y)}{\text{exposure}}) = a + bx$$

(25)

Ένα χαρακτηριστικό της κατανομής Poisson είναι ότι η μέση τιμή της είναι ίση με την διακύμανση. Κάποιες φορές όμως παρατηρούμε ότι η διακύμανση είναι μεγαλύτερη από την μέση τιμή. Κάτι τέτοιο είναι γνωστό με τον όρο υπερδιασπορά (Berk and MacDonald 2007) και δηλώνει ότι το μοντέλο δεν είναι κατάλληλο για τα δεδομένα. Συχνά ο λόγος που εμφανίζεται η υπερδιασπορά είναι η παράλειψη κάποιων επεξηγηματικών μεταβλητών.

Άλλο ένα γνωστό πρόβλημα με την παλινδρόμηση Poisson είναι η παρουσία μεγάλου αριθμού μηδενικών. Αν εκτελούνται δυο διεργασίες, μια να αποφασίζει αν υπάρχουν μηδέν γεγονότα ή αν υπάρχουν, και μια άλλη Poisson διεργασία να υπολογίζει τον αριθμό των γεγονότων, τότε θα υπάρχουν παραπάνω μηδενικά από όσα μπορεί να προβλέψει μια παλινδρόμηση Poisson. Ένα παράδειγμα για αυτό θα ήταν η κατανομή των τσιγάρων που καπνίζονται μέσα σε μια ώρα από τα μέλη μιας ομάδας ατόμων, όπου στην ομάδα αυτή υπάρχουν και μη-καπνιστές.

Αυτή η μέθοδος δουλεύει με την υπόθεση ότι οι παρατηρούμενες συχνότητες των απλότυπων αντιπροσωπεύουν μια τυχαία μεταβλητή η οποία ακολουθεί την κατανομή Poisson. Έτσι μπορούμε να εκτελέσουμε αυτό το μοντέλο, βάζοντας σαν εξαρτημένη μεταβλητή τις συχνότητες, και σαν παράγοντα επιρροής (ανεξάρτητες μεταβλητές) τους απλότυπους, αλλά και την κατάσταση (case/control).

Ιστορικά αυτά τα λεγόμενα log-linear μοντέλα έχουν χρησιμοποιηθεί αρκετά νωρίς για ανάλυση απλότυπων, για παράδειγμα στον υπολογισμό του συνελεστή συσχέτισης (LD), (Weir and Wilson 1986) αλλά και σε μελέτες γενετικής συσχέτισης ασθενειών με απλότυπους (Tiret, Amouyel et al. 1991). Το μοντέλο αυτό μπορεί να διατυπωθεί με τον εξής τύπο:

$$\log(n_j) = \mu_0 + \beta_0 y_j + \sum_{j=2}^{r} a_j z_j + \sum_{j=2}^{r} \beta_j z_j y_j$$

(26)

Το οποίο είναι ένα μοντέλο για την περιγραφή των 2 x r πινάκων συνάφειας απλότυπων σε σχέση με την ασθένεια. Τα β_j είναι οι συντελεστές που αντιστοιχούν στην αλληλεπίδραση (interaction) απλότυπου-ασθένειας και είναι όμοια με αυτά της

λογιστικής και της πολυωνυμικής παλινδρόμησης. Εδώ βγαίνει εύκολα το συμπέρασμα πως οι συντελεστές a_j και β_j είναι όμοιοι και στα τρία μοντέλα.

Για να εκτελέσουμε αυτόν τον έλεγχο στο Stata, αφού εισάγουμε τα δεδομένα, πρέπει να πληκτρολογήσουμε την παρακάτω εντολή:

```
xi: poisson frequency i.case*i.haplotype_nr
```

Με το `xi` στην αρχή της εντολής δηλώνουμε ότι θέλουμε να χρησιμοποιήσουμε μια κατηγορική μεταβλητή η οποία θα έχει έναν δείκτη. Σε αυτή την περίπτωση θα είναι το γινόμενο των μεταβλητών `case` και `haplotype_nr` όπου βάζοντας μπροστά τους το `i.` δηλώνουμε στο Stata ότι είναι οι μεταβλητές μας. Πέρα από αυτά με την εντολή `poisson` δηλώνουμε πως θέλουμε να κάνουμε παλινδρόμηση Poisson με εξαρτημένη μεταβλητή την `frequency` και ανεξάρτητη την `i.case*i.haplotype_nr`.

Και τα αποτελέσματα εμφανίζονται στην εικόνα 30:

```
i.case            _Icase_0-1          (naturally coded; _Icase_0 omitted)
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)
i.case*i.hapl~r   _IcasXhap_#_#       (coded as above)

Iteration 0:   log likelihood = -6133.2732
Iteration 1:   log likelihood = -5823.1098
Iteration 2:   log likelihood = -5820.7762
Iteration 3:   log likelihood = -5820.7737
Iteration 4:   log likelihood = -5820.7737

Poisson regression                                Number of obs   =         84
                                                  LR chi2(13)     =   16592.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -5820.7737                       Pseudo R2       =     0.5877

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
_Ihaplotyp~2 |  -.2244389   .0261886    -8.57   0.000    -.2757676   -.1731102
_Ihaplotyp~3 |  -1.583283    .042288   -37.44   0.000    -1.666166    -1.5004
_Ihaplotyp~4 |  -1.873937   .0478406   -39.17   0.000    -1.967703   -1.780171
_Ihaplotyp~5 |  -3.015109   .0807204   -37.35   0.000    -3.173318    -2.8569
_Ihaplotyp~6 |  -4.225312   .1453889   -29.06   0.000    -4.510269   -3.940355
_Ihaplotyp~7 |  -3.053088   .0821962   -37.14   0.000    -3.214189   -2.891986
_IcasXhap_~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_IcasXhap_~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_IcasXhap_~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_IcasXhap_~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_IcasXhap_~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_IcasXhap_~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |   6.304753   .0174528   361.25   0.000     6.270547    6.33896
------------------------------------------------------------------------------
```

Εικόνα 30: Αποτελέσματα της παλινδρόμησης poisson

Τα τελικά αποτελέσματα που λαμβάνουμε υπόψη είναι οι μεταβλητές _IcasXhap_~i, i=2…7 οι οποίες φαίνονται στις γραμμές 8-13, στην στήλη frequency του πίνακα της εικόνας 30.

Όπως και στις προηγούμενες μεθόδους, μπορούμε να εμφανίσουμε τα αποτελέσματα με την στήλη των λόγων των πιθανοτήτων κάθε απλότυπου:

```
xi: poisson frequency i.case*i.haplotype_nr, irr
```

Όπου εδώ το irr σημαίνει incidence-rate ratio (σχετικός λόγος πιθανοτήτων), και έχουμε το αποτέλεσμα της εικόνας 31.

```
i.case            _Icase_0-1          (naturally coded; _Icase_0 omitted)
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)
i.case*i.hapl~r   _IcasXhap_#_#       (coded as above)

Iteration 0:   log likelihood = -6133.2732
Iteration 1:   log likelihood = -5823.1098
Iteration 2:   log likelihood = -5820.7762
Iteration 3:   log likelihood = -5820.7737
Iteration 4:   log likelihood = -5820.7737

Poisson regression                              Number of obs   =         84
                                                LR chi2(13)     =   16592.09
                                                Prob > chi2     =     0.0000
Log likelihood = -5820.7737                     Pseudo R2       =     0.5877

------------------------------------------------------------------------------
   frequency |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .7346939   .0197029   -11.50   0.000     .6970743    .7743437
_Ihaplotyp~2 |   .7989644   .0209238    -8.57   0.000     .7589893    .8410449
_Ihaplotyp~3 |      .2053   .0086817   -37.44   0.000     .1889702     .223041
_Ihaplotyp~4 |   .1535181   .0073444   -39.17   0.000     .1397776    .1686094
_Ihaplotyp~5 |   .0490405   .0039586   -37.35   0.000     .0418645    .0574466
_Ihaplotyp~6 |   .0146208   .0021257   -29.06   0.000     .0109955    .0194413
_Ihaplotyp~7 |   .0472129   .0038807   -37.14   0.000     .0401879    .0554659
_IcasXhap_~2 |    .843754   .0349465    -4.10   0.000     .7779664    .9151047
_IcasXhap_~3 |   .9148121   .0607533    -1.34   0.180     .8031617    1.041983
_IcasXhap_~4 |   .7372685   .0588261    -3.82   0.000     .6305347    .8620698
_IcasXhap_~5 |   .4311594   .0702383    -5.16   0.000     .3133087    .5933396
_IcasXhap_~6 |   .9357639   .2130904    -0.29   0.771     .5988696    1.462178
_IcasXhap_~7 |    .922043    .119135    -0.63   0.530     .7157633    1.187772
------------------------------------------------------------------------------
```

Εικόνα 31: Αποτελέσματα της παλινδρόμησης poisson με incidence-rate ratio

Για την εκτέλεση του ελέγχου ισότητας των συντελεστών ισχύει η εντολή που χρησιμοποιούμε στην λογιστική παλινδρόμηση, απλά την εφαρμόζουμε στην μεταβλητή _IcasXhap_1_i, i=2….7.

```
testparm _IcasXhap_1_*
```

Όπου το σύμβολο ‘*’ δηλώνει πως θέλουμε όλα τα περιεχόμενα της μεταβλητής _casXhap_1_.

Και το αποτέλεσμα είναι το αναμενόμενο (εικόνα 32):

```
( 1)  [frequency]_IcasXhap_1_2 = 0
( 2)  [frequency]_IcasXhap_1_3 = 0
( 3)  [frequency]_IcasXhap_1_4 = 0
( 4)  [frequency]_IcasXhap_1_5 = 0
( 5)  [frequency]_IcasXhap_1_6 = 0
( 6)  [frequency]_IcasXhap_1_7 = 0

           chi2(  6) =   47.57
         Prob > chi2 =    0.0000
```

Εικόνα 32: Αποτελέσματα testparm για παλινδρόμηση poissson

Όπως και με τις άλλες δυο μεθόδους, μπορούμε να κάνουμε αυτόν τον έλεγχο για διάφορους συνδιασμούς των συντελεστών, για παράδειγμα:

```
test _IcasXhap_1_3 _IcasXhap_1_7
```

με αποτέλεσμα την εικόνα 33.

```
( 1)  [frequency]_IcasXhap_1_3 = 0
( 2)  [frequency]_IcasXhap_1_7 = 0

           chi2(  2) =    2.07
         Prob > chi2 =    0.3560
```

Εικόνα 33: Αποτέλεσμα του test για δυο μεταβλητές

Φυσικά μπορούμε να κάνουμε αυτόν τον έλεγχο με όσες από τις μεταβλητές που έχουν αντίστοιχο συντελεστή παλινδρόμησης επιθυμούμε, πχ:

```
test _IcasXhap_1_3 _IcasXhap_1_6 _IcasXhap_1_7
test _IcasXhap_1_2 _IcasXhap_1_4 _IcasXhap_1_5 _IcasXhap_1_7
```

Άσχετα με το ποια είναι η στρατηγική δειγματοληψίας που δημιούργησε τα δεδομένα μας, είναι ευρέως γνωστό ότι τα αποτελέσματα της προσαρμογής αυτών των μοντέλων στα δεδομένα είναι σχεδόν όμοια (Agresti 2002). Επομένως η επιλογή του μοντέλου πρέπει να βασίζεται σε παράγοντες που έχουν να κάνουν με την αληθοφάνεια και την ερμηνευτικότητα των αποτελεσμάτων. Να μπορούν δηλαδή να αναλυθούν εύκολα, και να παράγουν χρήσιμα και ξεκάθαρα αποτελέσματα. Επίσης ένας παράγοντας που πρέπει να λάβουμε υπόψη είναι το πόσο εύκαμπτο μπορεί να

είναι το μοντέλο και να επιτρέπει αλλαγές στα δεδομένα ή τις παραμέτρους του. Έχει αποδειχτεί ότι οι υπολογισμοί μέγιστης πιθανοφάνειας που αποκτούμε από τα μοντέλα με την μέθοδο της αναμενόμενης πιθανοφάνειας (prospective likelihood) είναι ίδιοι με αυτούς που αποκτώνται με την μέθοδο της αναδρομικής πιθανοφάνειας (retrospective likelihood) (Prentice and Pyke 1979; Chen 2003). Η ισότητα της λογιστικής παλινδρόμησης και των μοντέλων Poisson έχουν επίσης χρησιμοποιηθεί στο παρελθόν για την δημιουργία μεθόδων για τον εντοπισμό της αλληλεπίδρασης γονιδίων σε βιολογικό επίπεδο, (Umbach and Weinberg 1997).

Οι παραπάνω μέθοδοι είναι απλές προεκτάσεις των γενικευμένων γραμμικών μοντέλων και υποθέτουν πως α) Η απλοτυπική πιθανότητα ακολουθεί ένα αθροιστικό μοντέλο, β) Ότι η φάση του απλότυπου είναι γνωστή (βλέπε εισαγωγή – το πρόβλημα εύρεσης του απλότυπου (Lin, Cutler et al. 2002)) και γ) ότι ο πληθυσμός είναι σε ισορροπία Hardy-Weinberg. Το γενετικό μοντέλο της κληρονομικότητας μπορεί να λειτουργήσει με την χρήση των «απλο-γονότυπων» αντί για τους γονότυπους στην ανάλυση. Αυτό πραγματοποιείται εύκολα με τις μεθόδους που παρουσιάστηκαν χρησιμοποιώντας όλους τους συνδυασμούς των απλότυπων (h_1h_1, h_1h_2, κλπ). Ωστόσο στις μελέτες συσχέτισης με ασθενείς και μάρτυρες που χρησιμοποιούμε άτομα χωρίς συγγένεια, για να αντιμετωπίσουμε το πρόβλημα της εύρεσης του απλότυπου, χρησιμοποιούμε στατιστικές μεθόδους, συνήθως με χρήση του αλγορίθμου EM, ή κάποιον παρεμφερή (Niu 2004; Xu, Wu et al. 2004; Marchini, Cutler et al. 2006). Το να χρησιμοποιούμε τους απλότυπους σαν γνωστούς ίσως αποβεί προβληματικό (Lin and Huang 2007), επειδή όταν χρησιμοποιούμε πιθανοτικές μέθοδους, ο πιο πιθανός απλότυπος, δεν είναι απαραίτητα και ο σωστός.

Για την αντιμετώπιση αυτών των προβλημάτων έχουν αναπτυχθεί διάφορες μέθοδοι κατά καιρούς, για παράδειγμα η πρόσθεση κάποιου βάρους στους απλότυπους σύμφωνα με την πιθανότητα εμφάνισής τους (Zaykin, Westfall et al. 2002; French, Lumley et al. 2006). Η διαδικασία της παλινδρόμησης δεν αλλάζει, απλά προσθέτουμε επιπλέον μια στήλη με όνομα `prob` στον πίνακα 4, η οποία περιέχει τις αντίστοιχες πιθανότητες για κάθε απλότυπο.

Για παράδειγμα μετά την εισαγωγή των δεδομένων στο Stata εκτελούμε τις παρακάτω εντολές:

```
expand frequency
```

Αυτή η εντολή επεκτείνει τα δεδομένα έτσι ώστε να μπορέσουμε να εφαρμόσουμε στην συνέχεια την λογιστική παλινδρόμηση αλλά με πιθανοτικά βάρη. Αυτό γίνεται με την εντολή:

```
xi: logit case i.haplotype_nr [pweight=prob]
```

Με το `[pweight=prob]`να δηλώνει πως θα χρησιμοποιήσουμε πιθανοτικά βάρη, θα λάβουμε δηλαδή υπόψιν την αλληλεπίδραση των απλότυπων με βάση την πιθανότητα που έχουν να είναι αληθείς. Οι τιμές των συντελεστών παλινδρόμησης βλέπουμε πως έχουν επηρεαστεί, αν συγκρίνουμε τις εικόνες 18 και 34.

```
i.haplotype_nr    _Ihaplotype_1-7     (naturally coded; _Ihaplotype_1 omitted)

(sum of wgt is   9.3215e+03)
Iteration 0:   log pseudolikelihood = -8371.4204
Iteration 1:   log pseudolikelihood = -8354.2923
Iteration 2:   log pseudolikelihood = -8354.2368
Iteration 3:   log pseudolikelihood = -8354.2367

Logistic regression                               Number of obs   =      12406
                                                  Wald chi2(6)    =      47.56
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -8354.2367                 Pseudo R2       =     0.0021

------------------------------------------------------------------------------
             |               Robust
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.1698943   .0414196    -4.10   0.000    -.2510752   -.0887135
_Ihaplotyp~3 |  -.0890366   .0664134    -1.34   0.180    -.2192044    .0411312
_Ihaplotyp~4 |  -.3048031   .0797925    -3.82   0.000    -.4611935   -.1484128
_Ihaplotyp~5 |  -.8412774   .1629121    -5.16   0.000    -1.160579   -.5219756
_Ihaplotyp~6 |  -.0773611   .2239085    -0.35   0.730    -.5162136    .3614914
_Ihaplotyp~7 |  -.0781156   .1287041    -0.61   0.544     -.330371    .1741399
       _cons |  -.3083014   .0268189   -11.50   0.000    -.3608654   -.2557374
------------------------------------------------------------------------------
```

Εικόνα 34: Αποτελέσματα της λογιστικής παλινδρόμησης με πιθανοτικά βάρη

Ένας άλλος τρόπος θα ήταν η περίπτωση της εντολής `hapipf` η οποία δέχεται γονοτυπικά δεδομένα και υπολογίζει αυτόματα την απλοτυπική φάση (Mander 2001).

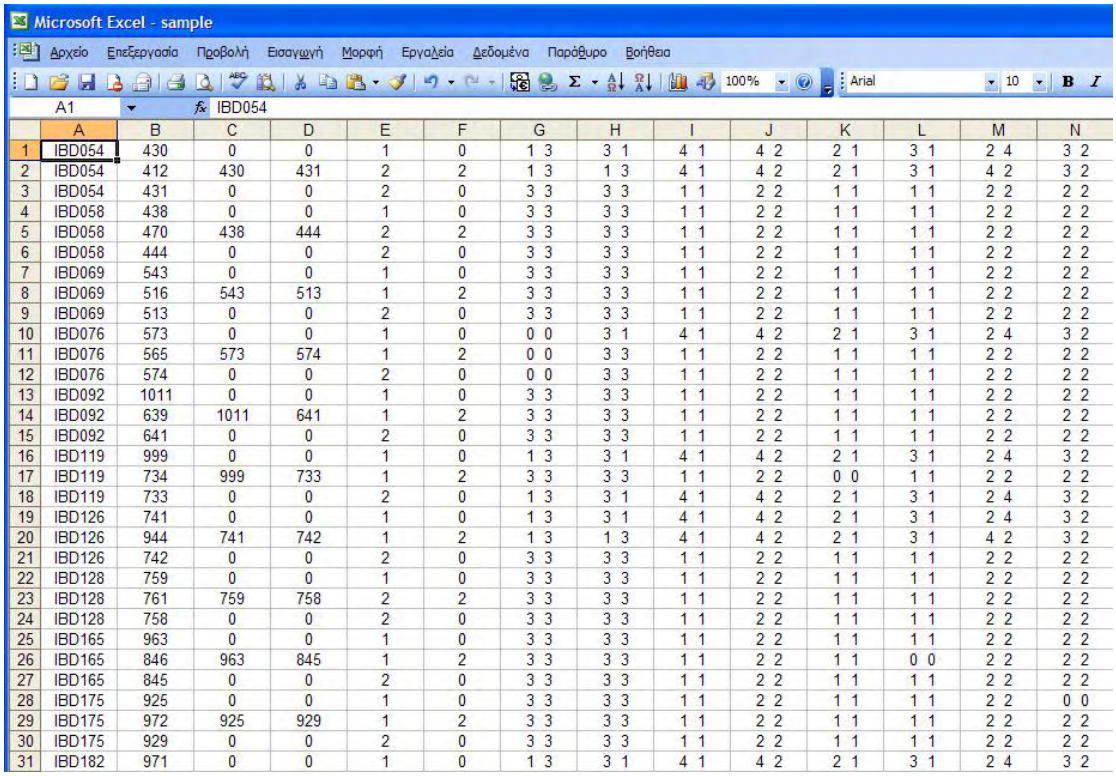
Επίσης έχουν αναπτυχθεί μέθοδοι τύπου score για αναμενόμενη πιθανότητα (Schaid, Rowland et al. 2002) και για αναδρομική πιθανότητα (Epstein and Satten 2003), καθώς και μέθοδους που αφορούν την αλληλεπίδραση των γονιδίων (Chen and Li 2008). Μια σύγκριση των μεθόδων έδειξε ότι τα αποτελέσματα είναι δύσκολο να συγκριθούν όταν η επιρροή του απλότυπου στην πιθανότητα της ασθένειας ακολουθεί ένα πολυπαραγοντικό μοντέλο. Ωστόσο για κυριαρχικά και υποχωρητικά μοντέλα, η

μέθοδος αναδρομικής πιθανότητας έχει αυξημένη αποτελεσματικότητα, χωρίς να παραμελούμε την αναμενόμενη πιθανότητα (Satten and Epstein 2004).

Επίσης γραφικά μοντέλα έχουν προταθεί από τον Thomas (Thomas 2005), τα οποία επιδιώκουν την ανάλυση γονοτυπικών δεδομένων αντί για απλοειδή, και λογαριθμικά-γραμμικά μοντέλα από τον Baker (Baker 2005) ο οποίος συμπεριλαμβάνει κάποιους περιβαλοντικούς συντελεστές οι οποίοι συμβάλουν στο αποτέλεσμα και λαμβάνει επίσης υπόψη τυχόν αποκλίσεις από την ισορροπία Hardy-Weinberg. Ο Lin και συνεργάτες έχουν επεκτείνει τις μεθόδους που παρουσιάστηκαν παραπάνω, με το να συμπεριλάβουν διάφορα είδη δειγματοληψίας σε ένα ενοποιημένο πλαίσιο εργασίας (Lin, Zeng et al. 2005). Εφάρμοσαν τις μεθόδους τους σε μια μελέτη για τον καρκίνο του μαστού η οποία έδειξε κάποιες σημαντικές επιπλοκές του καπνίσματος σε απλότυπους σχετικά με τον καρκίνο αυτού του τύπου.

2.4.6 Η εφαρμογή Haploview

Η εφαρμογή Haploview είναι ένα συχνά χρησιμοποιούμενο εργαλείο σε θέματα βιοπληροφορικής, το οποίο έχει δημιουργηθεί για την ανάλυση και την οπτικοποίηση ακολουθιών, αλλά και υπολογισμό της ανισορροπίας σύνδεσης (linkage disequilibrium) σε γενετικά δεδομένα. Επίσης είναι κατάλληλο για μελέτες γενετικής συσχέτισης, για την διαλογή κάποιων πολυμορφισμών και τέλος για τον υπολογισμό της συχνότητας των απλότυπων. Η εφαρμογή δημιουργήθηκε και εξελίσσεται από το εργαστήριο του Dr.Mark Daly στο ινστιτούτο του MIT/Harvard Broad Institute (http://www.broadinstitute.org).

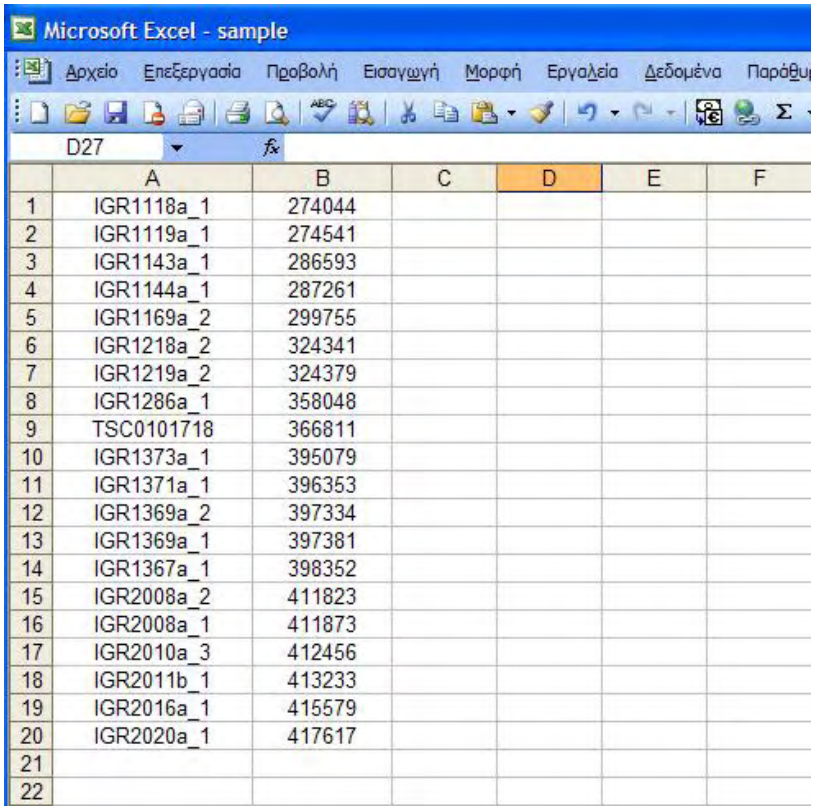


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	IBD054	430	0	0	1	0	1 3	3 1	4 1	4 2	2 1	3 1	2 4	3 2
2	IBD054	412	430	431	2	2	1 3	1 3	4 1	4 2	2 1	3 1	4 2	3 2
3	IBD054	431	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
4	IBD058	438	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
5	IBD058	470	438	444	2	2	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
6	IBD058	444	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
7	IBD069	543	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
8	IBD069	516	543	513	1	2	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
9	IBD069	513	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
10	IBD076	573	0	0	1	0	0 0	3 1	4 1	4 2	2 1	3 1	2 4	3 2
11	IBD076	565	573	574	1	2	0 0	3 3	1 1	2 2	1 1	1 1	2 2	2 2
12	IBD076	574	0	0	2	0	0 0	3 3	1 1	2 2	1 1	1 1	2 2	2 2
13	IBD092	1011	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
14	IBD092	639	1011	641	1	2	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
15	IBD092	641	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
16	IBD119	999	0	0	1	0	1 3	3 1	4 1	4 2	2 1	3 1	2 4	3 2
17	IBD119	734	999	733	1	2	3 3	3 3	1 1	2 2	0 0	1 1	2 2	2 2
18	IBD119	733	0	0	2	0	1 3	3 1	4 1	4 2	2 1	3 1	2 4	3 2
19	IBD126	741	0	0	1	0	1 3	3 1	4 1	4 2	2 1	3 1	2 4	3 2
20	IBD126	944	741	742	1	2	1 3	1 3	4 1	4 2	2 1	3 1	4 2	3 2
21	IBD126	742	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
22	IBD128	759	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
23	IBD128	761	759	758	2	2	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
24	IBD128	758	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
25	IBD165	963	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
26	IBD165	846	963	845	1	2	3 3	3 3	1 1	2 2	1 1	0 0	2 2	2 2
27	IBD165	845	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
28	IBD175	925	0	0	1	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	0 0
29	IBD175	972	925	929	1	2	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
30	IBD175	929	0	0	2	0	3 3	3 3	1 1	2 2	1 1	1 1	2 2	2 2
31	IBD182	971	0	0	1	0	1 3	3 1	4 1	4 2	2 1	3 1	2 4	3 2

Εικόνα 35: Τα δεδομένα του linkage format

Το Haploview μέχρι τώρα υποστηρίζει τις παρακάτω διαδικασίες: Μπορεί να υπολογίσει απλότυπους και ανισορροπία σύνδεσης με γονοτυπικά δεδομένα. Μπορεί να υπολογίσει την συχνότητα των απλότυπων στον πληθυσμό. Είναι δυνατή η εκτέλεση μελέτης γενετικής συσχέτισης με απλότυπους ή πολυμορφισμούς. Περιέχει υλοποίηση του αλγορίθμου επιλογής πολυμορφισμών ως δεικτών από τον Paul de Bakker. Υπάρχει η δυνατότητα λήψης δεδομένων από το HapMap, και τέλος μπορεί να απεικονίσει γραφικά τα αποτελέσματα όπως για παράδειγμα τα LD plots (Barrett 2009).

Στο πλαίσιο αυτής της εργασίας, έγινε εισαγωγή δεδομένων τα οποία παρέχονται ήδη από την εφαρμογή, και ακολουθεί μια αναφορά των αποτελεσμάτων. Τα δεδομένα γίνονται δεκτά σε πέντε συνολικά μορφές (linkage, Haps, HapMap, PHASE, PLINK) αλλά εμάς μας ενδιέφερε περισσότερο η πρώτη (linkage format) η οποία μπορούσε να χρησιμοποιηθεί και για μελέτες ασθενών - μαρτύρων. Τα αρχεία αυτών των δεδομένων είναι αρχεία .ped, και απαιτείται η χρήση ενός προγράμματος που ονομάζεται Makeped, για να αποκτήσουν την σωστή μορφή, και να είναι αποδεκτά από το Haploview. Επίσης υπάρχει ένα συνοδευτικό αρχείο το οποίο περιέχει πληροφορίες για την θέση κάθε πολυμορφισμού πάνω στο εκάστοτε χρωμόσωμα.



	A	B	C	D	E	F
1	IGR1118a_1	274044				
2	IGR1119a_1	274541				
3	IGR1143a_1	286593				
4	IGR1144a_1	287261				
5	IGR1169a_2	299755				
6	IGR1218a_2	324341				
7	IGR1219a_2	324379				
8	IGR1286a_1	358048				
9	TSC0101718	366811				
10	IGR1373a_1	395079				
11	IGR1371a_1	396353				
12	IGR1369a_2	397334				
13	IGR1369a_1	397381				
14	IGR1367a_1	398352				
15	IGR2008a_2	411823				
16	IGR2008a_1	411873				
17	IGR2010a_3	412456				
18	IGR2011b_1	413233				
19	IGR2016a_1	415579				
20	IGR2020a_1	417617				
21						
22						

Εικόνα 36: Οι πολυμορφισμοί και η θέση τους

Στην εικόνα 35 βλέπουμε τα περιεχόμενα του linkage format. Στις πρώτες έξι στήλες καταγράφουμε κάποια δεδομένα που αφορούν τα άτομα. Στην πρώτη στήλη έχουμε ένα μοναδικό αλφαριθμητικό κωδικό που προσδιορίζει την οικογένεια που ανήκει το άτομο. Στην δεύτερη στήλη έχουμε έναν μοναδικό κωδικό που προσδιορίζει ένα άτομο. Στην τρίτη και τέταρτη στήλη έχουμε τους κωδικούς του πατέρα και της μητέρας του ατόμου αντίστοιχα. Στην πέμπτη στήλη έχουμε το φύλο

(1=αρσενικό, 2=θυλικό), ενώ στην έκτη στήλη έχουμε την κατάσταση επίδρασης. Η κατάσταση επίδρασης δηλώνει αν το άτομο αυτό επηρεάζεται από κάποιον παράγοντα που μελετάμε (0=άγνωστο, 1=δεν επηρεάζεται, 2=επηρεάζεται).

Από την έβδομη στήλη και μετά γίνεται η καταγραφή των γονότυπων. Στην εικόνα 35 φαίνονται με αριθμούς (1-4), και η αντιστοιχία είναι 1=A, 2=C, 3=G και 4=T. Ο χρήστης έχει τη δυνατότητα να εισάγει τα δεδομένα και με τους χαρακτήρες των βάσεων, δηλαδή A T G και C.

Το αρχείο με τα δεδομένα αυτά, πρέπει να συνοδεύεται και από ένα αρχείο info, το οποίο περιέχει μια λίστα με τους πολυμορφισμούς που ερευνούμε (εικόνα 36), καθώς και την θέση τους στο χρωμόσωμα. Η θέση των πολυμορφισμών είναι ο αριθμός των βάσεων μετρώντας από την αρχή του χρωμοσώματος.

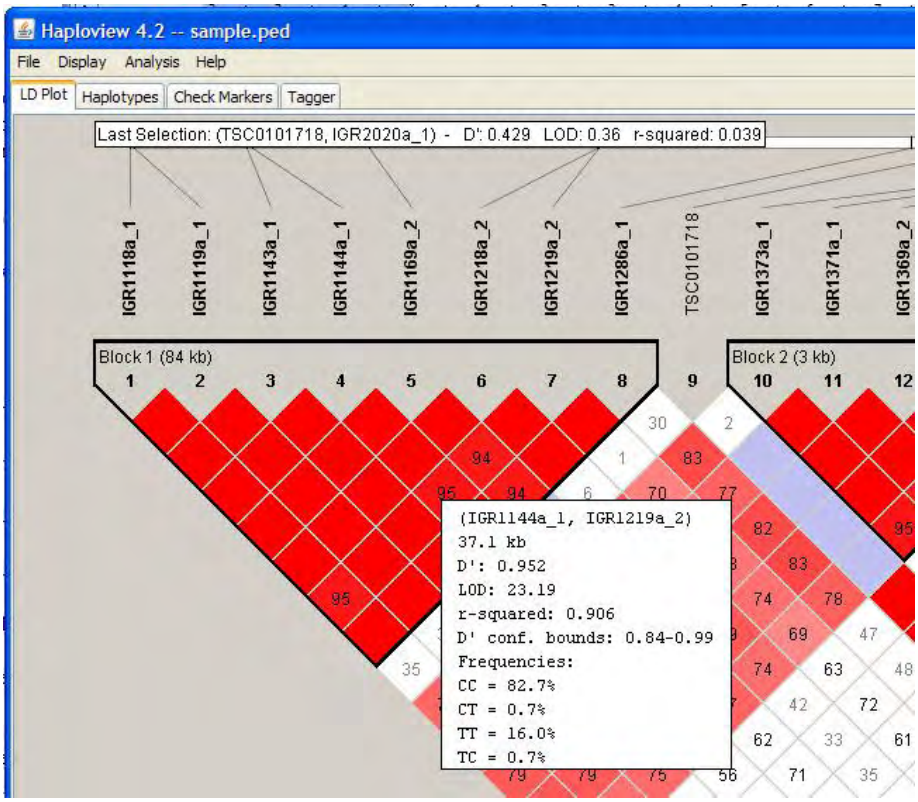
Εικόνα 37: LD plot

Αφού εισάγουμε τα δεδομένα, η εφαρμογή βγάζει κάποια αποτελέσματα. Το πρώτο είναι ένα διάγραμμα που δείχνει την συσχέτιση των πολυμορφισμών, και το πόσο ισχυρή είναι (εικόνα 37). Οι πιο ισχυρά σχετιζόμενοι πολυμορφισμοί ομαδοποιούνται και σχηματίζουν απλότυπους. Ο χρήστης έχει την δυνατότητα να επιλέξει έναν από τους διάφορους τρόπους που υπάρχουν για την επιλογή της ομάδας

των πολυμορφισμών που θα αποτελέσουν έναν απλότυπο, δεδομένου της ισχυρής μεταξύ τους συσχέτισης (Gabriel, Schaffner et al. 2002; Wang, Akey et al. 2002).

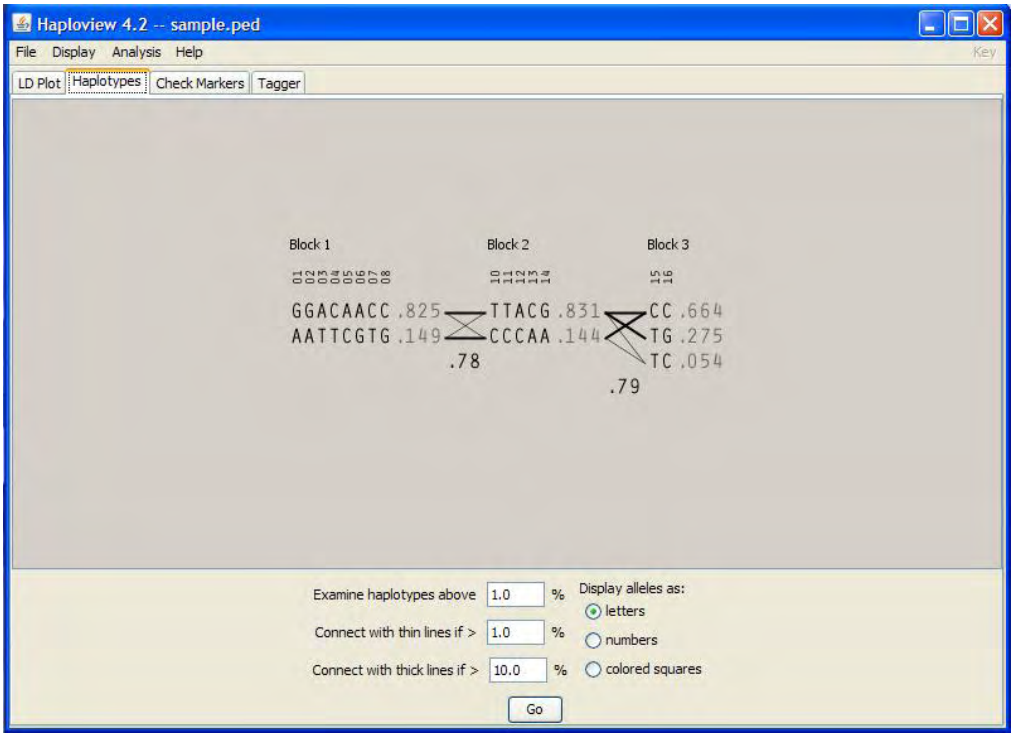
Επίσης ο χρήστης μπορεί να επιλέξει χειροκίνητα άλλες περιοχές για ανάλυση και όχι απαραίτητα αυτές που επιλέγει αυτόματα το Haploview. Μετά την επιλογή των περιοχών συσχέτισης, η εφαρμογή δημιουργεί τους απλότυπους και υπολογίζει τις συχνότητές τους στον πληθυσμό. Αυτή η απεικόνιση φαίνεται στην εικόνα 39. Οι γραμμές απεικονίζουν την μετάβαση από μια περιοχή ισχυρής συσχέτισης στην επόμενη, και το πάχος της γραμμής είναι ανάλογο με την συχνότητα μετάβασης σε αυτήν την περιοχή. Η γραμμή απεικονίζει επίσης τον βαθμό συσχέτισης (LD) D΄ μεταξύ των δυο περιοχών, χρησιμοποιώντας τον απλότυπο σαν το αλληλόμορφο για αυτήν την περιοχή. Διάφορες παράμετροι είναι στην επιλογή του χρήστη, όπως τα αλληλόμορφα να απεικονίζονται με γράμματα ή αριθμούς ή χρωματιστά κουτάκια, αλλά και η απεικόνιση μόνο των απλότυπων που ξεπερνούν ένα ρυθμιζόμενο κατώφλι (ή όριο) του πληθυσμού.

Κάθε τετραγωνάκι στο σχήμα απεικονίζει την συσχέτιση μεταξύ δύο κάθε φορά πολυμορφισμών. Το χρώμα υποδεικνύει το πόσο μεγάλη είναι η συσχέτιση, για παράδειγμα το κόκκινο χρώμα σημαίνει πολύ ισχυρή συσχέτιση. Η εφαρμογή υπολογίζει τους συντελεστές συσχέτισης D΄ και r^2 καθώς και τις συχνότητες κάθε περίπτωσης βάσεων όπως φαίνεται στην εικόνα 38.

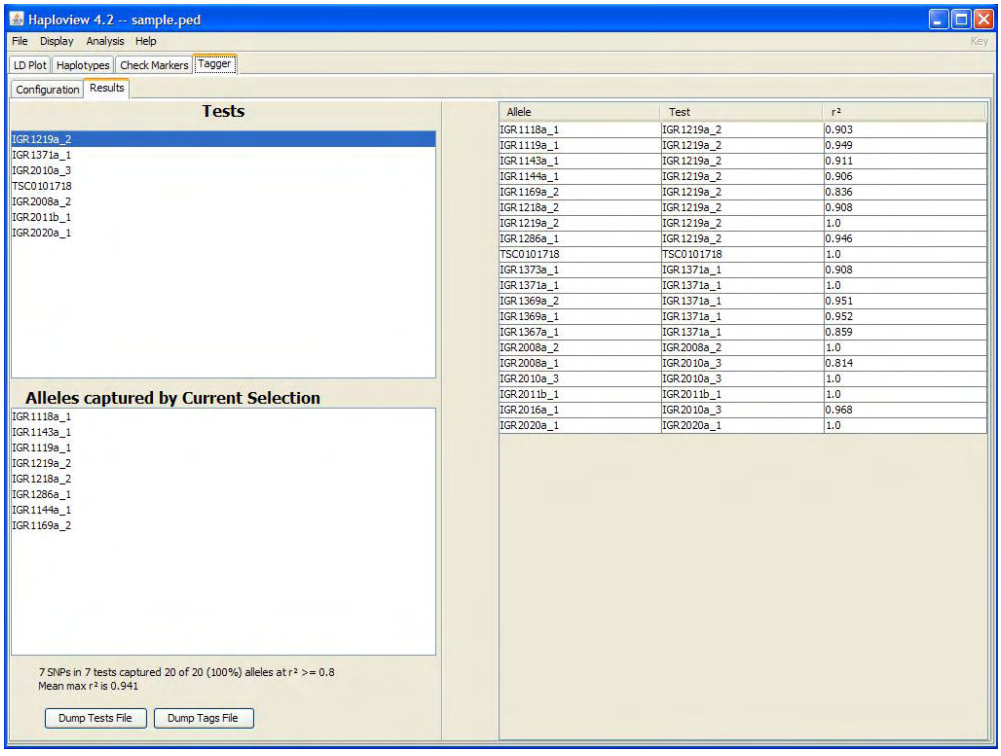


Εικόνα 38: Τα στοιχεία που αντιστοιχούν σε ένα ζευγάρι πολυμορφισμών

Το τελευταίο κομμάτι των αποτελεσμάτων αποτελείται από μια λίστα ζευγαριών πολυμορφισμών που εξετάσαμε. Κάθε πολυμορφισμός εξετάστηκε ως προς την συσχέτισή του με τους υπόλοιπους, με αποτέλεσμα μια λίστα που φαίνεται στην εικόνα 40. Για κάθε ζευγάρι υπολογίστηκε ο συντελεστής συσχέτισης r^2, και η εφαρμογή μπορεί να εμφανίσει όλους τους πολυμορφισμούς οι οποίοι έχουν συντελεστή συσχέτισης πάνω από κάποιο όριο, το οποίο επίσης μπορεί να επιλεχθεί από τον χρήστη. Συνοπτικά έχουμε μια αναφορά που περιέχει n πολυμορφισμούς, και nxn υπολογισμούς συσχέτισης (γίνεται έλεγχος και με ζευγάρι δυο ίδιων πολυμορφισμών από το Haploview).



Εικόνα 39: Αποτελέσματα απλότυπων



Εικόνα 40: Η λίστα με τις συσχετίσεις των πολυμορφισμών

2.5 Μελέτες με μεταβλητή συνεχούς τιμής

Εκτός από τις μελέτες με ασθενείς και μάρτυρες, ασχοληθήκαμε και με ένα άλλο είδος μελετών που πραγματοποιούν απλοτυπική ανάλυση. Σε αυτή την περίπτωση έχουμε έναν παράγοντα κινδύνου, και ερευνούμε την σχέση του με τον εκάστοτε απλότυπο. Ψάχνουμε δηλαδή μια συσχέτιση μεταξύ απλότυπου και παράγοντα κινδύνου. Αυτό βέβαια δεν έχει μεγάλη διαφορά από τις μελέτες ασθενών και μαρτύρων, η μόνη διαφορά είναι ότι ο παράγοντας είναι μεταβλητή που δεν παίρνει δυο τιμές για κάθε άτομο. Είναι δηλαδή αριθμητική μεταβλητή και όχι κατηγορική. Χαρακτηριστικό παράδειγμα τέτοιου παράγοντα είναι οι τιμές της χοληστερόλης στο αίμα (Lu, Inazu et al. 2003). Συνεπώς θα έχουμε μια λίστα ανθρώπων οι οποίοι θα ανήκουν σε κάποια απλοτυπική ομάδα, και θα έχουν ο καθένας την δικιά του τιμή για την χοληστερόλη.

2.5.1 Απλή γραμμική παλινδρόμηση

Η απλή γραμμική παλινδρόμηση είναι ένας στατιστικός έλεγχος που εξετάζει την σχέση μεταξύ δυο μεταβλητών X και Y. Η κύρια διαδικασία είναι ο υπολογισμός των ελάχιστων τετραγώνων των μεταβλητών, από μια ευθεία γραμμή η οποία περνάει μέσα από τις παρατηρήσεις στην γραφική παράσταση. Αυτή η ευθεία είναι τοποθετημένη με τέτοιο τρόπο, ώστε να ελαχιστοποιεί το άθροισμα των κάθετων αποστάσεων των στοιχείων από αυτήν (Eberly 2007). Το επίθετο «απλή» παλινδρόμηση αναφέρεται στο γεγονός ότι αυτή η παλινδρόμηση είναι η πιο απλή στην στατιστική. Όπως είδαμε υπάρχουν και άλλες μέθοδοι παλινδρόμησης.

Η ευθεία που αναφέρεται παραπάνω έχει κλίση ίση με την συσχέτιση των δυο μεταβλητών X και Y διορθωμένη από τον λόγο των τυπικών αποκλίσεων των μεταβλητών αυτών.

Αν υποθέσουμε ότι τα δεδομένα μας έχουν μέγεθος n με $\{Y_i, X_i\}$ όπου i=1,2,…..,n., τότε ο σκοπός μας είναι να βρούμε την εξίσωση της ευθείας:

$$y = \alpha + \beta x \tag{27}$$

η οποία θα ταιριάζει καλύτερα στα δεδομένα. Όταν λέμε ότι θα ταιριάζει καλύτερα αναφερόμαστε στην προσέγγιση των ελάχιστων τετραγώνων, έτσι ώστε να έχουμε το

μικρότερο άθροισμα (εικόνα 41). Αν συμβαίνει αυτό, τότε οι αριθμοί α και β λύνουν το ακόλουθο πρόβλημα ελαχιστοποίησης:

$$\min_{\alpha,\beta} Q(\alpha,\beta)\text{, όπου } Q(\alpha,\beta)=\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

(28)

Με την χρήση μαθηματικής ανάλυσης ή γεωμετρίας αποδεικνύεται ότι οι τιμές των α και β που ελαχιστοποιούν την συνάρτηση Q είναι οι εξής:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{j=1}^{n}y_j/n}{\sum_{i=1}^{n}(x_i^2)-(\sum_{i=1}^{n}x_i)^2/n}$$

$$=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\overline{x^2}-\bar{x}^2}=\frac{Cov[x,y]}{Var[x]}=r_{xy}\frac{s_y}{s_x}$$

(29)

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$$

(30)

όπου r_{xy} είναι ο συντελεστής συσχέτισης του δείγματος μεταξύ του X και Y, s_x η τυπική απόκλιση του X και s_y αντίστοιχα η τυπική απόκλιση του Y.

Αν αντικαταστήσουμε τα $\hat{\alpha}$ και $\hat{\beta}$ στην εξίσωση 30 προκύπτει η ακόλουθη μορφή:

$$\frac{y-\bar{y}}{s_y}=r_{xy}\frac{x-\bar{x}}{s_x}$$

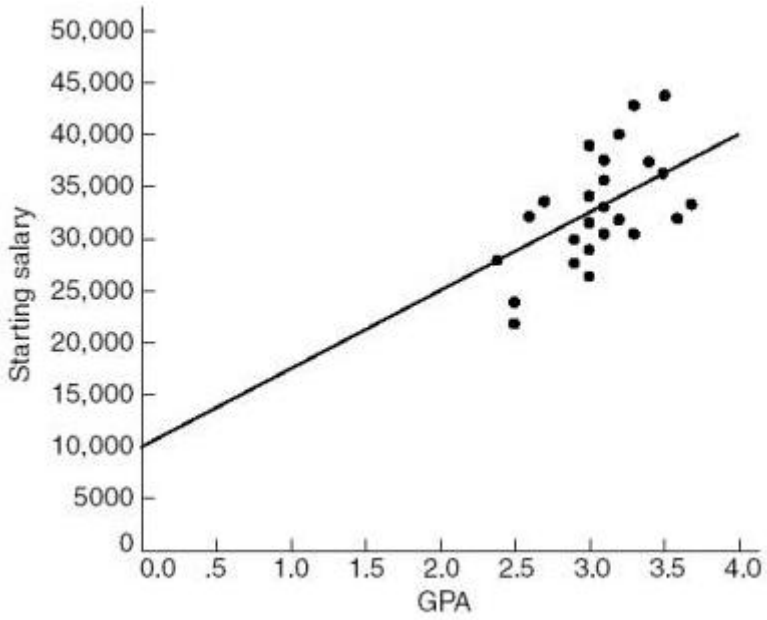
(31)

Αυτό δείχνει το ρόλο που παίζει ο συντελεστής συσχέτισης στην ευθεία της παλινδρόμησης.

- Η ευθεία της παλινδρόμησης περνάει από το κέντρο μάζας $(\overline{x}, \overline{y})$.
- Το σύνολο των στοιχείων είναι μηδέν, εάν το μοντέλο περιλαμβάνει σφάλμα:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

- Τα $\hat{\alpha}$ και $\hat{\beta}$ είναι αμερόληπτα, αυτό σημαίνει ότι ερμηνεύουμε το μοντέλο στοχαστικά, πράγμα που σημαίνει ότι πρέπει να υποθέσουμε πως για κάθε τιμή του x, η αντίστοιχη τιμή του y παράγεται σαν ένα μέσο αποτέλεσμα $\alpha + \beta x$ με μια επιπλέον μεταβλητή σφάλματος ε. Αυτό το σφάλμα πρέπει να είναι κατά μέσο όρο μηδέν, για κάθε τιμή του x. Σύμφωνα με αυτήν την ερμηνεία οι εκτιμήτριες ελάχιστων τετραγώνων $\hat{\alpha}$ και $\hat{\beta}$ θα είναι τυχαίες μεταβλητές, και θα εκτιμούν αμερόληπτα τις πραγματικές τιμές των α και β.



Εικόνα 41: Απλή παλινδρόμηση που δείχνει την συσχέτιση μεταξύ πρώτου μισθού και μέσου όρου βαθμολογίας φοιτητών

Στην περίπτωση των απλοτύπων χρησιμοποιούμε την πολλαπλή παλινδρόμηση με τον ακόλουθο μαθηματικό τύπο να προσαρμόζεται στα δεδομένα:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_k \mathbf{x}_k \quad \textbf{(32)}$$

για k απλότυπους.

Τα δεδομένα που παίρνουμε για την εκτέλεση της παλινδρόμησης έχουν γενικά την μορφή του πίνακα 6.

Απλότυπος	Συχνότητα	Συνεχής μεταβλητή (πχ τιμή χοληστερόλης στο αίμα)	Διακύμανση
απλότυπ_1	96	50.1	1.4
απλότυπ_2	50	55.1	3.2
…	43	53.2	3.4
…	37	50.3	3.6
απλότυπ_n	19	50.1	5.4

Πίνακας 6: Γενική μορφή των δεδομένων

Τις περισσότερες φορές, η διακύμανση δεν μας δίνεται από τις μελέτες και συνεπώς πρέπει να την υπολογίσουμε.
Για τον υπολογισμό αυτό χρησιμοποιούμε τον τύπο:

$$\text{var} = sd^2 / n$$

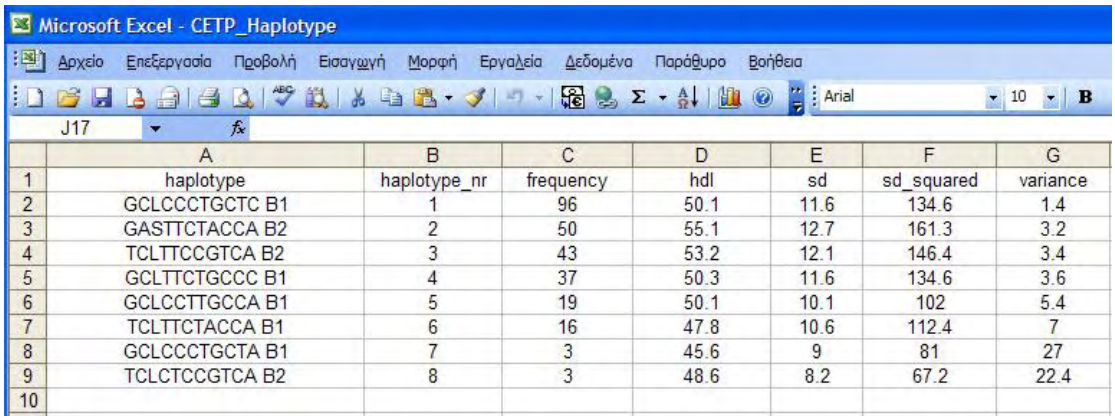
(33)

Αλλά επίσης χρήσιμος είναι και ο τύπος της τυπικής απόκλισης:

$$se^2 = sd^2 / n$$

(34)

και έτσι προκύπτουν άλλες δυο στήλες στον πίνακα 6 μια για την τυπική απόκλιση (sd) και μια για το sd^2 όπως φαίνεται στην εικόνα 42.



	A	B	C	D	E	F	G
1	haplotype	haplotype_nr	frequency	hdl	sd	sd_squared	variance
2	GCLCCCTGCTC B1	1	96	50.1	11.6	134.6	1.4
3	GASTTCTACCA B2	2	50	55.1	12.7	161.3	3.2
4	TCLTTCCGTCA B2	3	43	53.2	12.1	146.4	3.4
5	GCLTTCTGCCC B1	4	37	50.3	11.6	134.6	3.6
6	GCLCCTTGCCA B1	5	19	50.1	10.1	102	5.4
7	TCLTTCTACCA B1	6	16	47.8	10.6	112.4	7
8	GCLCCCTGCTA B1	7	3	45.6	9	81	27
9	TCLCTCCGTCA B2	8	3	48.6	8.2	67.2	22.4
10							

Εικόνα 42: Τα δεδομένα στο excel για την απλή παλινδρόμηση

Το Stata έχει την εντολή `regress` για την απλή παλινδρόμηση, αλλά σε αυτήν την εργασία χρησιμοποιούμε την διακύμανση σαν επιπλέον παράγοντα που επηρεάζει το αποτέλεσμα. Γι' αυτό το λόγο χρησιμοποιείται η εντολή `vwls` (variance-weighted least-squares regression). Για να εκτελέσουμε αυτόν τον έλεγχο πληκτρολογούμε τις παρακάτω εντολές:

```
gen se=sqrt(variance)
```

Με την εντολή `gen` δημιουργούμε μια μεταβλητή `se` (standard error) η οποία θα χρειαστεί αργότερα και είναι η τετραγωνική ρίζα της διακύμανσης, όπως προκύπτει από τις σχέσεις 32 και 33.

```
xi: vwls hdl i.haplotype_nr,sd(se)
```

Με το `xi` στην αρχή της εντολής δηλώνουμε ότι θέλουμε να χρησιμοποιήσουμε μια κατηγορική μεταβλητή η οποία θα έχει έναν δείκτη. Με την εντολή `vwls` δηλώνουμε πως θέλουμε να κάνουμε vwls παλινδρόμηση με εξαρτημένη μεταβλητή την `hdl` και ανεξάρτητη την `i.haplotype_nr`.
Χρησιμοποιούμε έμμεσα την διακύμανση βάζοντας σαν επιλογή μετά την εντολή το `sd(se)`, αντί για την χρήση του `fweight` ή του `aweight` που είδαμε στις παλινδρομήσεις των μελετών ασθενών και μαρτύρων. Μετά την εκτέλεση της εντολής, έχουμε τα αποτελέσματα της εικόνας 43.

```
i.haplotype_nr    _Ihaplotype_1-8     (naturally coded; _Ihaplotype_1 omitted)

Variance-weighted least-squares regression          Number of obs   =         8
Goodness-of-fit chi2(0)     =       .               Model chi2(7)   =     10.33
Prob > chi2                 =       .               Prob > chi2     =    0.1707
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |          5   2.144761     2.33   0.020      .7963456    9.203654
_Ihaplotyp~3 |   3.100002    2.19089     1.41   0.157     -1.194064    7.394068
_Ihaplotyp~4 |   .2000008   2.236068     0.09   0.929     -4.182612    4.582613
_Ihaplotyp~5 |   7.76e-16   2.607681     0.00   1.000     -5.110961    5.110961
_Ihaplotyp~6 |  -2.299999   2.898275    -0.79   0.427     -7.980514    3.380516
_Ihaplotyp~7 |        -4.5   5.329165    -0.84   0.398     -14.94497    5.944971
_Ihaplotyp~8 |        -1.5   4.878524    -0.31   0.758     -11.06173    8.061732
       _cons |        50.1   1.183216    42.34   0.000      47.78094    52.41906
------------------------------------------------------------------------------
```

Εικόνα 43: Αποτελέσματα της vwls παλινδρόμησης

Όσον αφορά την ερμηνεία των αποτελεσμάτων, ισχύουν και εδώ οι βασικές αρχές της παλινδρόμησης. Κάθε απλότυπος έχει έναν συντελεστή που δηλώνει την ένταση της επιρροής στο συνολικό μοντέλο. Με το p-value βλέπουμε αν ισχύει η μηδενική υπόθεση ή όχι. Σε αυτήν την περίπτωση η εξαρτημένη μεταβλητή συνεχούς τιμής είναι τα επίπεδα της χοληστερόλης στο αίμα. Οι ανεξάρτητες μεταβλητές που επηρεάζουν τις τιμές της χοληστερόλης είναι οι διάφοροι απλότυποι, που σε αυτήν την έρευνα είναι οκτώ διαφορετικοί.

Η μέθοδος αυτή της vwls παλινδρόμησης μπορεί να μας πει, ότι οι τιμές της εξαρτημένης μεταβλητής επηρεάζονται από κάποιον απλότυπο. Όταν έχουμε θετικό συντελεστή παλινδρόμησης αυτό σημαίνει πως ο απλότυπος αυτός συμβάλει στην αύξηση των τιμών HDL-c ή CETP αντίστοιχα. Ενώ αντίθετα ένα αρνητικό πρόσημο σημαίνει πως ο απλότυπος αυτός ρίχνει τις αντίστοιχες τιμές. Η τελευταία τιμή από τους συντελεστές παλινδρόμησης την εικόνα 43 μας δείχνει την τιμή της HDL που αντιστοιχεί στον πιο κοινό απλότυπο του δείγματος, δηλαδή τον πρώτο. Οι συντελεστές των υπόλοιπων απλότυπων δηλώνουν την απόκλιση από αυτή την τιμή. Για παράδειγμα η τιμή της HDL που αντιστοιχεί στον δεύτερο απλότυπο είναι 50.1, αν προσθέσουμε τον αντίστοιχο συντελεστή 5 θα έχουμε ως αποτέλεσμα την τιμή που αντιστοιχεί στα δεδομένα, δηλαδή 55.1 (πίνακας 7).

ΚΕΦΑΛΑΙΟ 3 – ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

3.1 ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1.1 Μελέτες απλότυπων με τις τιμές της HDL χοληστερίνης (συνεχής μεταβλητή)

Συνολικά έγιναν τρεις παλινδρομήσεις, μια για κάθε μελέτη. Οι δυο από αυτές τις παλινδρομήσεις εμφάνισαν στατιστικά σημαντικά αποτελέσματα σύμφωνα με τον X^2 του Pearson που είδαμε παραπάνω (2.4.1), με p-value 0.0011 και 0.0032. Η παλινδρόμηση με το μικρότερο p-value είχε και τον υψηλότερο X^2=25.84. Στο σύνολο μελετήθηκαν 22 απλότυποι σε σχέση με τις τιμές της HDL-χοληστερόλης αλλά και 5 απλότυποι σε σχέση με την συγκέντρωση της CETP (cholesterol ester transfer protein), μιας πρωτείνης η οποία μεταφέρει τα μόρια της χοληστερόλης στο αίμα.

Η πρώτη μελέτη περιέχει απλοτυπική ανάλυση της περιοχής του υποκινητή του γονιδίου της CET πρωτείνης, και έχει τα δεδομένα που ψάχνουμε (Lu, Inazu et al. 2003). Αυτά αποτελούνται από οκτώ απλότυπους (εικόνα 44), τρείς από αυτούς επιδρούν αρνητικά στις τιμές της HDL-χοληστερόλης, ενώ από τους υπόλοιπους οι δυο έχουν αρκετά υψηλό συντελεστή παλινδρόμησης δείχνοντας ότι παίζουν σημαντικό ρόλο. Για τον τέταρτο απλότυπο ο συντελεστής είναι κοντά στο μηδέν, πράγμα που σημαίνει πως δεν επηρεάζει τόσο την HDL.

Στατιστικά σημαντική επιρροή θεωρούμε όμως μόνο αυτήν του δεύτερου απλότυπου, καθώς μόνο εκεί βλέπουμε p-value μικρότερο από την τιμή 0.05 και μπορούμε να απορρίψουμε την υπόθεση πως δεν υπάρχει σχέση με την ασθένεια. Γενικά βλέπουμε επίσης ότι για τον X^2 στην μελέτη αυτή, το p-value είναι επίσης μεγαλύτερο από την προαναφερθείσα τιμή και συνεπώς είναι ένα στοιχείο που προδιαθέτει αρνητικά για την αξιοπιστία των αποτελεσμάτων. Τα δεδομένα που πήραμε από αυτήν την μελέτη φαίνονται στον πίνακα 7.

haplotype	haplotype_nr	frequency	hdl	sd	sd_squared	variance
GCLCCCTGCTC B1	1	96	50.1	11.6	134.6	1.4
GASTTCTACCA B2	2	50	55.1	12.7	161.3	3.2
TCLTTCCGTCA B2	3	43	53.2	12.1	146.4	3.4
GCLTTCTGCCC B1	4	37	50.3	11.6	134.6	3.6
GCLCCTTGCCA B1	5	19	50.1	10.1	102	5.4
TCLTTCTACCA B1	6	16	47.8	10.6	112.4	7
GCLCCCTGCTA B1	7	3	45.6	9	81	27
TCLCTCCGTCA B2	8	3	48.6	8.2	67.2	22.4

Πίνακας 7: Απλοτυπικά δεδομένα για τις τιμές της hdl (mg/dl)

```
i.haplotype_nr    _Ihaplotype_1-8     (naturally coded; _Ihaplotype_1 omitted)

Variance-weighted least-squares regression          Number of obs   =         8
Goodness-of-fit chi2(0)     =         .             Model chi2(7)   =     10.33
Prob > chi2                 =         .             Prob > chi2     =    0.1707
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |          5   2.144761     2.33   0.020      .7963456    9.203654
_Ihaplotyp~3 |   3.100002    2.19089     1.41   0.157     -1.194064    7.394068
_Ihaplotyp~4 |   .2000008   2.236068     0.09   0.929     -4.182612    4.582613
_Ihaplotyp~5 |   7.76e-16   2.607681     0.00   1.000     -5.110961    5.110961
_Ihaplotyp~6 |  -2.299999   2.898275    -0.79   0.427     -7.980514    3.380516
_Ihaplotyp~7 |       -4.5   5.329165    -0.84   0.398     -14.94497    5.944971
_Ihaplotyp~8 |       -1.5   4.878524    -0.31   0.758     -11.06173    8.061732
       _cons |        50.1   1.183216    42.34   0.000      47.78094    52.41906
------------------------------------------------------------------------------
```

Εικόνα 44: Η πρώτη και λιγότερο σημαντική παλινδρόμηση (Lu, Inazu et al. 2003)

Η δεύτερη μελέτη η οποία αφορά την απλοτυπική ανάλυση του γονιδίου της CETP, είναι η πιο αξιόπιστη από πλευράς αποτελεσμάτων, με p-value < 0.05 και X^2=25.84. Βλέπουμε αρκετούς απλότυπους να συμβάλουν σημαντικά στις τιμές της HDL, και επιπλέον με θετικούς συντελεστές παλινδρόμησης (εικόνα 45). Οι πρώτοι τρείς απλότυποι συμβάλουν με υψηλούς συντελεστές, ο απλότυπος 2 με συντελεστή 7.8, ο απλότυπος 3 με 11.3 και ο τέταρτος με συντελεστή 10.1. Τα δεδομένα αυτής της μελέτης φαίνονται στον πίνακα 8.

haplotype	haplotype_nr	frequency	hdl	sd	sd_squared	variance
a1	1	80	47.2	14	196	2.5
c1	2	45	55	14	196	4.4
c2	3	35	58.5	14	196	5.6
a2	4	13	57.3	14	196	15
b2	5	9	55.7	14	196	22
d1	6	6	56.5	14	196	32.6
e1	7	6	64.3	14	196	32.6
d2	8	4	56.1	14	196	49
b1	9	2	54.6	14	196	98

Πίνακας 8: Απλοτυπικά δεδομένα για τις τιμές της hdl (mg/dl)

```
i.haplotype_nr    _Ihaplotype_1-9     (naturally coded; _Ihaplotype_1 omitted)

Variance-weighted least-squares regression          Number of obs   =         9
Goodness-of-fit chi2(0)     =         .             Model chi2(8)   =     25.84
Prob > chi2                 =         .             Prob > chi2     =    0.0011
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   7.799999   2.626785     2.97   0.003     2.651595     12.9484
_Ihaplotyp~3 |       11.3    2.84605     3.97   0.000     5.721844    16.87815
_Ihaplotyp~4 |       10.1     4.1833     2.41   0.016     1.900881    18.29912
_Ihaplotyp~5 |        8.5   4.949748     1.72   0.086    -1.201327    18.20133
_Ihaplotyp~6 |   9.299999   5.924525     1.57   0.116    -2.311857    20.91186
_Ihaplotyp~7 |       17.1   5.924525     2.89   0.004     5.488146    28.71186
_Ihaplotyp~8 |   8.899998    7.17635     1.24   0.215     -5.16539    22.96539
_Ihaplotyp~9 |   7.399998   10.02497     0.74   0.460    -12.24858    27.04858
       _cons |       47.2   1.581139    29.85   0.000     44.10103    50.29898
------------------------------------------------------------------------------
```

Εικόνα 45: Η παλινδρόμηση της ισχυρότερης σε στατιστική σημαντικότητα μελέτης (Bauerfeind, Knoblauch et al. 2002)

Επιπλέον την μεγαλύτερη επιρροή φαίνεται να έχει ο απλότυπος 7, ο οποίος με p-value = 0.004 έχει τον υψηλότερο συντελεστή παλινδρόμησης (17.1) από όλους και θεωρούμε ότι έχει αρκετά μεγάλη επιρροή στις τιμές της HDL. Όσον αφορά τους υπόλοιπους απλότυπους, παρουσιάζουν επίσης υψηλούς συντελεστές παλινδρόμησης, αλλά δεν είναι στατιστικά σημαντικοί καθώς το p-value είναι πάνω από το όριο που έχουμε θέσει στα πλαίσια του διαστήματος εμπιστοσύνης 95%.

Στην Τρίτη και τελευταία μελέτη η οποία αφορά επίσης απλοτυπική ανάλυση στο γονίδιο της CETP, έχουμε πέντε απλότυπους και ερευνούμε την συσχέτισή τους με τις τιμές της HDL (εικόνα 46), αλλά και την συγκέντρωση της πρωτείνης CETP (εικόνα 47). Όσον αφορά την HDL, βλέπουμε γενικά πολύ μικρούς συντελεστές παλινδρόμησης, από τους οποίους δυο είναι στατιστικά σημαντικοί. Οστόσο η επιρροή τους είναι ουδέτερη, με συντελεστή παλινδρόμησης κοντά στο μηδέν δεν έχουμε ούτε θετική ούτε αρνητική επιρροή. Τα δεδομένα όσον αφορά τις τιμές της HDL φαίνονται στον πίνακα 9.

haplotype	haplotype_nr	frequency	hdl	se	variance
1	1	290	0.9	0.007	0.000049
2	2	87	0.9	0.011	0.000121
3	3	32	0.92	0.013	0.000169
4	4	50	0.93	0.012	0.000144
5	5	197	0.94	0.009	0.000081

Πίνακας 9: Απλοτυπικά δεδομένα για τις τιμές της hdl (mmol/l)

```
i.haplotype_nr    _Ihaplotype_1-5     (naturally coded; _Ihaplotype_1 omitted)

Variance-weighted least-squares regression          Number of obs    =         5
Goodness-of-fit chi2(0)     =         .             Model chi2(4)    =     15.88
Prob > chi2                 =         .             Prob > chi2      =    0.0032
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -1.66e-18   .0130384    -0.00   1.000    -.0255548     .0255548
_Ihaplotyp~3 |         .02   .0147648     1.35   0.176    -.0089385     .0489386
_Ihaplotyp~4 |         .03   .0138924     2.16   0.031     .0027713     .0572287
_Ihaplotyp~5 |         .04   .0114018     3.51   0.000      .017653      .062347
       _cons |          .9       .007   128.57   0.000     .8862802     .9137197
------------------------------------------------------------------------------
```

Εικόνα 46: Η παλινδρόμηση της μελέτης με το μεγαλύτερο δείγμα (HDL-c) (Klerkx, Tanck et al. 2003)

Επίσης βλέπουμε στον δεύτερο απλότυπο έναν πολύ μεγάλο συντελεστή (-1.66e-18) με το e-18 να σημαίνει μετακίνηση της υποδιαστολής κατά 18 θέσεις προς τα δεξιά) και p-value πολύ μεγαλύτερο από το αποδεκτό όριο. Αυτό ίσως οφείλεται στην μεγάλη διαφορά συχνοτήτων μεταξύ του απλότυπου 2 και του 1.

Η αντίστοιχη παλινδρόμηση για τις τιμές της CETP έδειξε υψηλότερο X^2 από αυτήν της HDL. Οι συντελεστές παλινδρόμησης είναι όλοι κοντά στο μηδέν, και από αυτούς μόνο η επιρροή που αντιστοιχεί στον τέταρτο και πέμπτο απλότυπο είναι στατιστικά σημαντική. Η επιρροή αυτή είναι της τάξεως του -0.07 για τον απλότυπο 4 και -0.15 περίπου για τον απλότυπο 5. Σε συνδυασμό με την τιμή της CETP η οποία έιναι χαμηλότερη για τον απλότυπο αυτό, μπορούμε να συμπεράνουμε μια μικρή επιρροή, όχι όμως και για τις αντίστοιχες τιμές της HDL, όπου δεν υπάρχει σχεδόν καθόλου. Τα δεδομένα όσον αφορά τις τιμές της CETP φαίνονται στον πίνακα 10.

haplotype	haplotype_nr	frequency	CETP	se	variance
1	1	290	1.96	0.018	0.000324
2	2	87	2	0.030	0.0009
3	3	32	1.92	0.035	0.001225
4	4	50	1.89	0.036	0.001296
5	5	197	1.81	0.023	0.000529

Πίνακας 10: Απλοτυπικά δεδομένα για τις τιμές της CETP (mg/ml)

```
i.haplotype_nr    _Ihaplotype_1-5     (naturally coded; _Ihaplotype_1 omitted)

Variance-weighted least-squares regression          Number of obs    =         5
Goodness-of-fit chi2(0)     =         .             Model chi2(4)    =     35.57
Prob > chi2                 =         .             Prob > chi2      =    0.0000
------------------------------------------------------------------------------
        cetp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |         .04   .0349857     1.14   0.253    -.0285708    .1085707
_Ihaplotyp~3 |  -.0400001   .0393573    -1.02   0.309     -.117139    .0371389
_Ihaplotyp~4 |  -.0700001   .0402492    -1.74   0.082    -.1488871     .008887
_Ihaplotyp~5 |  -.1500001   .0292062    -5.14   0.000    -.2072431   -.0927571
       _cons |       1.96        .018   108.89   0.000     1.924721    1.995279
------------------------------------------------------------------------------
```

Εικόνα 47: Η παλινδρόμηση της μελέτης με το μεγαλύτερο δείγμα (CETP) (Klerkx, Tanck et al. 2003)

3.1.2 Μελέτες με ασθενείς και μάρτυρες (case-control)

Συνολικά πήραμε δεδομένα απλότυπων από 19 σχετικές μελέτες ασθενών-μαρτύρων οι οποίες αναφέρονται στην βιβλιογραφία. Για κάθε μελέτη εκτελέσαμε και τις τρείς μεθόδους που περιγράψαμε παραπάνω (λογιστική, πολυωνυμική και Poisson παλινδρόμηση), αν και σχεδόν ισοδύναμες (Agresti 2002) διαφέρουν αρκετά στην μέθοδο υλοποίησης και τον υπολογισμό των συντελεστών. Σε κάποιες μελέτες υπήρχαν δεδομένα απλότυπων για περισσότερα από ένα γονίδια, ή δυο και παραπάνω σύνολα απλότυπων, συνεπώς ο αριθμός των παλινδρομήσεων είναι μεγαλύτερος από τον αναμενόμενο αριθμό.

Παρατηρήσαμε 8 μελέτες που είχαν p-value<0.05 στον έλεγχο X^2 πράγμα που κάνει τα αποτελέσματά τους στατιστικώς σημαντικά (Παράρτημα 1). Οι αντίστοιχοι X^2 είχαν αρκετά μεγάλο εύρος τιμών το οποίο βρίσκεται μεταξύ 76 και 9. Επίσης υπήρχε μεγάλη ποικιλία περιπτώσεων στα αποτελέσματα των παλινδρομήσεων. Για παράδειγμα είδαμε μελέτη με στατιστική σημαντικότητα, αλλά κανέναν απλότυπο να διαφοροποιείται από το σύνολο (Zee, Cook et al. 2005). Οι περισσότερες μελέτες που είχαν p-value>0.05 στον έλεγχο X^2 δεν είχαν ιδιαίτερα υψηλούς συντελεστές παλινδρόμησης και κατ' επέκταση δεν φανέρωσαν κάποια συσχέτιση με την ασθένεια. Μια μελέτη η οποία είχε p-value λίγο πάνω από την περιοχή απόρριψης της μηδενικής υπόθεσης, δηλαδή λίγο παραπάνω από 0.05

εμφάνισε κάποια συσχέτιση με ασθένεια σε δυο απλότυπους (Kankova, Stejskalova et al. 2007). Τέλος υπήρξαν και μελέτες οι οποίες έδειξαν την συσχέτιση ενός και μόνο απλότυπου με την ασθένεια, καθώς μόνο για αυτόν τον απλότυπο η συσχέτιση ήταν στατιστικά σημαντική.

Ακολουθεί η περαιτέρω ανάλυση κάποιων χαρακτηριστικών αποτελεσμάτων.

```
Logistic regression                                     Number of obs   =      12401
                                                        LR chi2(6)      =      50.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -8317.8958                             Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_Ihaplotyp~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_Ihaplotyp~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_Ihaplotyp~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_Ihaplotyp~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_Ihaplotyp~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =      12401
                                                        LR chi2(6)      =      50.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -8317.8958                             Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_Ihaplotyp~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_Ihaplotyp~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_Ihaplotyp~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_Ihaplotyp~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_Ihaplotyp~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         84
                                                        LR chi2(13)     =   16592.09
                                                        Prob > chi2     =     0.0000
Log likelihood = -5820.7737                             Pseudo R2       =     0.5877

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.3083014   .0268178   -11.50   0.000    -.3608632   -.2557395
_Ihaplotyp~2 |  -.2244389   .0261886    -8.57   0.000    -.2757676   -.1731102
_Ihaplotyp~3 |  -1.583283    .042288   -37.44   0.000    -1.666166    -1.5004
_Ihaplotyp~4 |  -1.873937   .0478406   -39.17   0.000    -1.967703   -1.780171
_Ihaplotyp~5 |  -3.015109   .0807204   -37.35   0.000    -3.173318    -2.8569
_Ihaplotyp~6 |  -4.225312   .1453889   -29.06   0.000    -4.510269   -3.940355
_Ihaplotyp~7 |  -3.053088   .0821962   -37.14   0.000    -3.214189   -2.891986
_IcasXhap_~2 |  -.1698943   .0414179    -4.10   0.000    -.2510719   -.0887167
_IcasXhap_~3 |  -.0890366   .0664107    -1.34   0.180    -.2191992    .0411259
_IcasXhap_~4 |  -.3048031   .0797892    -3.82   0.000    -.4611872   -.1484191
_IcasXhap_~5 |  -.8412774   .1629055    -5.16   0.000    -1.160566   -.5219884
_IcasXhap_~6 |  -.0663921   .2277182    -0.29   0.771    -.5127115    .3799273
_IcasXhap_~7 |  -.0811634   .1292076    -0.63   0.530    -.3344057    .1720789
       _cons |   6.304753   .0174528   361.25   0.000     6.270547     6.33896
------------------------------------------------------------------------------
```

Εικόνα 48: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Haplotypes extending across ACE are associated with Alzheimer's disease" (Kehoe, Katzov et al. 2003)

Η πρώτη μελέτη που αναλύθηκε αφορούσε τη σχέση των απλότυπων του γονιδίου ACE με την ασθένεια του Alzheimer. Στα αποτελέσματα που βλέπουμε στην εικόνα 48, έχουμε μια καλή περίπτωση συσχέτισης. Παρατηρούμε υψηλό X^2 με p-value<0.05, πράγμα που σημαίνει αρκετά αξιόπιστα αποτελέσματα. Μπορούμε να διακρίνουμε στατιστική σημαντικότητα για τρείς απλότυπους συγκεκριμένα, τους 2, 4 και 5. Παρατηρούμε πως οι συντελεστές τους έχουν αρνητικό πρόσημο. Αυτό σημαίνει πως έχουν αρνητική επίδραση στην ασθένεια, μειώνουν δηλαδή τις πιθανότητες να εμφανιστεί αυτή η ασθένεια στο μέλλον. Ένα τμήμα των δεδομένων που χρησιμοποιήσαμε φαίνεται στον πίνακα 11.

haplotype	frequency	haplotype_nr	case	frequency_percent
AAT	202	1	1	0,5
GTC	136	2	1	0,33
GAT	37	3	1	0,09
GAC	16	4	1	0,04
ATC	4	5	1	0,01
GTT	2	6	1	0,01
AAC	11	7	1	0,03
AAT	150	1	0	0,45
GTC	130	2	0	0,39
GAT	26	3	0	0,08
GAC	19	4	0	0,06
ATC	5	5	0	0,01
GTT	0	6	0	0
AAC	6	7	0	0,02

Πίνακας 11: Απλοτυπικά δεδομένα ασθενών-μαρτύρων

Για να δούμε καλύτερα τι συμβαίνει επαναλαμβάνουμε την πολυωνιμική παλινδρόμηση, και ζητάμε τους συντελεστές με την μορφή του σχετικού λόγου πιθανοτήτων (εικόνα 49).

```
Multinomial logistic regression                 Number of obs   =      12401
                                                LR chi2(6)      =      50.22
                                                Prob > chi2     =     0.0000
Log likelihood = -8317.8958                     Pseudo R2       =     0.0030

------------------------------------------------------------------------------
        case |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |    .843754   .0349465    -4.10   0.000     .7779664    .9151047
_Ihaplotyp~3 |   .9148121   .0607533    -1.34   0.180     .8031617    1.041983
_Ihaplotyp~4 |   .7372685   .0588261    -3.82   0.000     .6305347    .8620698
_Ihaplotyp~5 |   .4311594   .0702382    -5.16   0.000     .3133087    .5933396
_Ihaplotyp~6 |   .9357639   .2130904    -0.29   0.771     .5988696    1.462178
_Ihaplotyp~7 |    .922043    .119135    -0.63   0.530     .7157633    1.187772
------------------------------------------------------------------------------
(case==0 is the base outcome)
```

Εικόνα 49: Απεικόνιση των συντελεστών με Odds ratio – επέκταση της εικόνας 48

Βλέπουμε λοιπόν για τον δεύτερο απλότυπο ότι ο λόγος των πιθανοτήτων είναι περίπου 0.84. Αν είχαμε σαν αποτέλεσμα τον αριθμό 1, αυτό θα σήμαινε πως δεν υπάρχει καμιά επίδραση από αυτόν τον απλότυπο και ότι η πιθανότητα ασθένειας δεν αυξάνεται αλλά ούτε μειώνεται όπως έχουμε πει παραπάνω. Συνεπώς αν αφαιρέσουμε τον αριθμό 0.84 από την μονάδα θα έχουμε σαν αποτέλεσμα τον αριθμό: 0.16. Αυτός ο αριθμός αντιπροσωπεύει την αρνητική επιρροή του απλότυπου στην ασθένεια. Έτσι μπορούμε να πούμε ότι ο απλότυπος 2 μειώνει την πιθανότητα εμφάνισης αυτής της ασθένειας κατά 16%. Με το ίδιο σκεπτικό βλέπουμε ακόμη μεγαλύτερη επιρροή στον απλότυπο 4 κατά 27% και στον απλότυπο 5 κατά 57%.

Σε μια άλλη μελέτη για την σχέση των απλότυπων του γονιδίου ADAM33 με το άσθμα, είχαμε στα δεδομένα περιπτώσεις όπου είχαν συχνότητα μηδέν. Οι απλότυποι αυτοί υπήρχαν μόνο σε ασθενείς, αλλά δεν παρατηρήθηκαν σε κανέναν μάρτυρα (Kedda, Duffy et al. 2006). Έτσι τα δεδομένα αντιμετωπίστηκαν διαφορετικά στην περίπτωση της λογιστικής παλινδρόμησης, στην οποία για τους απλότυπους αυτούς δεν υπολογίστηκε συντελεστής και εξαιρέθηκαν από την διαδικασία. Αυτό είναι αναμενόμενο καθώς αν δεν υπάρχουν καθόλου άτομα από μια ομάδα, τότε αυτόματα βγάζουμε το συμπέρασμα ότι σίγουρα αυτοί οι απλότυποι παίζουν κάποιο ρόλο, ανάλογα την ομάδα που βρίσκονται (ασθενείς/μάρτυρες).

Με αυτόν τον τρόπο όμως, δεν μπορούμε να υπολογίσουμε αυτήν την σχέση ποσοτικά, καθώς δεν έχουμε ούτε p-value αλλά ούτε και συντελεστές παλινδρόμησης. Στη συνέχεια όμως παρατηρήσαμε ότι η πολυωνιμική και η Poisson παλινδρόμηση υπολόγισαν και τους απλότυπους αυτούς χάρη στις διαφορές υλοποίησης που έχουν (υπολογίζουν με διαφορετικό τρόπο τον σχετικό λόγο των πιθανοτήτων). Στην εικόνα 50 φαίνεται η λογιστική παλινδρόμηση και οι απλότυποι 14-15 παραλείπονται από την διαδικασία, σε αντίθεση με την εικόνα 51, καθώς δεν μπορούμε να υπολογίσουμε

τον σχετικό λόγο πιθανοτήτων με μηδενική συχνότητα σε εκείνη τη θέση. Τα απλοτυπικά δεδομένα που επεξεργαστήκαμε φαίνονται στον πίνακα 12. Παρατηρήστε τις συχνότητες (στήλη frequency) των απλότυπων 14 και 15 (στήλη haplotype_nr) για ασθενείς (case=1) και μάρτυρες (case=0).

haplotype	haplotype_nr	frequency	case
GCGAGCCCCA	1	190	1
GCGCGTCCCA	2	92	1
ACGAGCCCCA	3	73	1
GCGCGTCCGA	4	61	1
ATAAACACGA	5	49	1
ACGCATCCCA	6	24	1
GCGCGCCCCA	7	12	1
GCGAGTCCCA	8	18	1
ATGCACATCG	9	6	1
ATGCGTCCCA	10	6	1
ACGCGTCCGA	11	3	1
GCGCGCCCGT	12	6	1
GCAAGCACCA	13	6	1
ATGAACCCGA	14	6	1
ACGAGTCCCA	15	6	1
ACGCACCCCA	16	49	1
GCGAGCCCCA	1	161	0
GCGCGTCCCA	2	80	0
ACGAGCCCCA	3	47	0
GCGCGTCCGA	4	47	0
ATAAACACGA	5	43	0
ACGCATCCCA	6	24	0
GCGCGCCCCA	7	9	0
GCGAGTCCCA	8	4	0
ATGCACATCG	9	9	0
ATGCGTCCCA	10	4	0
ACGCGTCCGA	11	4	0
GCGCGCCCGT	12	2	0
GCAAGCACCA	13	1	0
ATGAACCCGA	14	0	0
ACGAGTCCCA	15	0	0
ACGCACCCCA	16	33	0

Πίνακας 12: Απλοτυπικά δεδομένα ασθενών-μαρτύρων

Logistic regression

Number of obs	= 1063
LR chi2(13)	= 16.33
Prob > chi2	= 0.2317
Log likelihood = -721.04525	Pseudo R2 = 0.0112

case	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ihaploty~_2	-.0258578	.1866652	-0.14	0.890	-.3917149	.3399993
_Ihaploty~_3	.2746921	.2155216	1.27	0.202	-.1477225	.6971067
_Ihaploty~_4	.0951066	.2216853	0.43	0.668	-.3393887	.5296018
_Ihaploty~_5	-.0349995	.2348155	-0.15	0.882	-.4952294	.4252304
_Ihaploty~_6	-.1656197	.3079085	-0.54	0.591	-.7691094	.4378699
_Ihaplotyp~7	.1220624	.4537827	0.27	0.788	-.7673355	1.01146
_Ihaplotyp~8	1.338458	.5630541	2.38	0.017	.234892	2.442023
_Ihaplotyp~9	-.5710848	.5378216	-1.06	0.288	-1.625196	.4830262
_Ihaploty~10	.2398454	.6543248	0.37	0.714	-1.042608	1.522299
_Ihaploty~11	-.4533018	.7712378	-0.59	0.557	-1.9649	1.058296
_Ihaploty~12	.9329926	.8234932	1.13	0.257	-.6810244	2.54701
_Ihaploty~13	1.62614	1.085422	1.50	0.134	-.5012483	3.753528
_Ihaploty~16	.229693	.2493703	0.92	0.357	-.2590637	.7184498
_cons	.1656197	.1071183	1.55	0.122	-.0443284	.3755678

Εικόνα 50: Λογιστική παλινδρόμηση στα δεδομένα της μελέτης "ADAM33 haplotypes are associated with asthma in a large Australian population" (Kedda, Duffy et al. 2006)

Η πολυωνιμική παλινδρόμηση ήταν η πιο βολική στην ερμηνεία των αποτελεσμάτων (εικόνα 52). Με υψηλό X^2 και p-value<0.05 έδειξε μεγάλη συσχέτιση στον απλότυπο 8 με σχετικό λόγο πιθανοτήτων 3.81. Κάτι τέτοιο σημαίνει πως όποιος έχει τον απλότυπο 8, έχει 3,81 φορές παραπάνω κίνδυνο να εμφανίσει άσθμα σε σχέση με τον πιο κοινό απλότυπο. Οι υπόλοιποι απλότυποι δεν έδειξαν κάποια συσχέτιση σύμφωνα με τις μεθόδους παλινδρόμησης, καθώς τα p-value τους δεν κατέβηκαν κάτω από 0.05 και έτσι δεν απορρίψαμε την μηδενική υπόθεση, ότι δηλαδή οι απλότυποι δεν σχετίζονται με την ασθένεια.

Multinomial logistic regression — Number of obs = 1075; LR chi2(15) = 30.15; Prob > chi2 = 0.0114; Log likelihood = -721.04525; Pseudo R2 = 0.0205

case	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
_Ihaploty~_2	-.0258578	.1866652	-0.14	0.890	-.3917149	.3399993
_Ihaploty~_3	.2746921	.2155216	1.27	0.202	-.1477225	.6971067
_Ihaploty~_4	.0951066	.2216853	0.43	0.668	-.3393887	.5296018
_Ihaploty~_5	-.0349995	.2348155	-0.15	0.882	-.4952294	.4252304
_Ihaploty~_6	-.1656197	.3079085	-0.54	0.591	-.7691094	.4378699
_Ihaplotyp~7	.1220624	.4537827	0.27	0.788	-.7673355	1.01146
_Ihaplotyp~8	1.338458	.5630541	2.38	0.017	.234892	2.442023
_Ihaplotyp~9	-.5710848	.5378216	-1.06	0.288	-1.625196	.4830262
_Ihaploty~10	.2398454	.6543248	0.37	0.714	-1.042608	1.522299
_Ihaploty~11	-.4533018	.7712378	-0.59	0.557	-1.9649	1.058296
_Ihaploty~12	.9329926	.8234932	1.13	0.257	-.6810244	2.54701
_Ihaploty~13	1.62614	1.085422	1.50	0.134	-.5012483	3.753528
_Ihaploty~14	36.07641	2.74e+07	0.00	1.000	-5.37e+07	5.37e+07
_Ihaploty~15	36.07641	2.74e+07	0.00	1.000	-5.37e+07	5.37e+07
_Ihaploty~16	.229693	.2493703	0.92	0.357	-.2590637	.7184498
_cons	.1656197	.1071183	1.55	0.122	-.0443284	.3755678

(case==0 is the base outcome)

Poisson regression — Number of obs = 32; LR chi2(31) = 1495.70; Prob > chi2 = 0.0000; Log likelihood = -69.518842; Pseudo R2 = 0.9149

frequency	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Icase_1	.1655652	.1071179	1.55	0.122	-.044382	.3755123
_Ihaploty~_2	-.6994155	.1367882	-5.11	0.000	-.9675155	-.4313156
_Ihaploty~_3	-1.231267	.1657922	-7.43	0.000	-1.556214	-.9063205
_Ihaploty~_4	-1.231267	.1657922	-7.43	0.000	-1.556214	-.9063205
_Ihaploty~_5	-1.320221	.1716577	-7.69	0.000	-1.656664	-.983778
_Ihaploty~_6	-1.903363	.2188076	-8.70	0.000	-2.332218	-1.474509
_Ihaplotyp~7	-2.884226	.3425251	-8.42	0.000	-3.555563	-2.212889
_Ihaplotyp~8	-3.69517	.5061793	-7.30	0.000	-4.687263	-2.703077
_Ihaplotyp~9	-2.884226	.3425251	-8.42	0.000	-3.555563	-2.212889
_Ihaploty~10	-3.69517	.5061793	-7.30	0.000	-4.687263	-2.703077
_Ihaploty~11	-3.69517	.5061793	-7.30	0.000	-4.687263	-2.703077
_Ihaploty~12	-4.388278	.7114805	-6.17	0.000	-5.782754	-2.993802
_Ihaploty~13	-5.081418	1.003091	-5.07	0.000	-7.047439	-3.115397
_Ihaploty~14	-20.05859	1787.501	-0.01	0.991	-3523.496	3483.378
_Ihaploty~15	-20.05859	1787.501	-0.01	0.991	-3523.496	3483.378
_Ihaploty~16	-1.584922	.1910856	-8.29	0.000	-1.959443	-1.210401
_IcasXhap~_2	-.0258219	.1866657	-0.14	0.890	-.39168	.3400362
_IcasXhap~_3	.274736	.2155198	1.27	0.202	-.147675	.6971471
_IcasXhap~_4	.095171	.2216827	0.43	0.668	-.3393192	.5296611
_IcasXhap~_5	-.0349357	.2348133	-0.15	0.882	-.4951613	.4252898
_IcasXhap~_6	-.1655652	.3079055	-0.54	0.591	-.7690489	.4379186
IcasXhap~7	.1221433	.4537828	0.27	0.788	-.7672547	1.011541
IcasXhap~8	1.338572	.5630581	2.38	0.017	.2349987	2.442146
IcasXhap~9	-.571045	.5378269	-1.06	0.288	-1.625166	.4830764
_IcasXhap~10	.2398994	.6543332	0.37	0.714	-1.04257	1.522369
_IcasXhap~11	-.4532388	.7712458	-0.59	0.557	-1.964853	1.058375
_IcasXhap~12	.9330075	.8234919	1.13	0.257	-.6810069	2.547022
_IcasXhap~13	1.626147	1.085415	1.50	0.134	-.5012264	3.753521
_IcasXhap~14	16.60331	1787.501	0.01	0.993	-3486.834	3520.04
_IcasXhap~15	16.60331	1787.501	0.01	0.993	-3486.834	3520.04
_IcasXhap~16	.2297649	.2493684	0.92	0.357	-.2589882	.7185181
_cons	5.081438	.0788097	64.48	0.000	4.926974	5.235903

Εικόνα 51: Πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "ADAM33 haplotypes are associated with asthma in a large Australian population" (Kedda, Duffy et al. 2006)

Multinomial logistic regression — Number of obs = 1075; LR chi2(15) = 30.15; Prob > chi2 = 0.0114; Log likelihood = -721.04525; Pseudo R2 = 0.0205

case	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
_Ihaploty~_2	.9744737	.1819003	-0.14	0.890	.6758968	1.404947
_Ihaploty~_3	1.316125	.2836535	1.27	0.202	.8626705	2.007935
_Ihaploty~_4	1.099776	.2438042	0.43	0.668	.7122056	1.698256
_Ihaploty~_5	.9656059	.2267392	-0.15	0.882	.6094311	1.529943
_Ihaploty~_6	.8473684	.260912	-0.54	0.591	.4634256	1.549403
_Ihaplotyp~7	1.129825	.5126949	0.27	0.788	.4642484	2.749613
_Ihaplotyp~8	3.813158	2.147014	2.38	0.017	1.264772	11.49628
_Ihaplotyp~9	.5649123	.303822	-1.06	0.288	.1968731	1.620972
_Ihaploty~10	1.271053	.8316813	0.37	0.714	.3525342	4.582747
_Ihaploty~11	.6355263	.4901419	-0.59	0.557	.1401699	2.881458
_Ihaploty~12	2.542105	2.093406	1.13	0.257	.5060983	12.76886
_Ihaploty~13	5.084211	5.518514	1.50	0.134	.605774	42.67136
_Ihaploty~14	4.65e+15	1.27e+23	0.00	1.000	0	.
_Ihaploty~15	4.65e+15	1.27e+23	0.00	1.000	0	.
_Ihaploty~16	1.258214	.3137611	0.92	0.357	.7717738	2.051251

(case==0 is the base outcome)

Εικόνα 52: Απεικόνιση των συντελεστών με Odds ratio – επέκταση της εικόνας 51

Άλλη περίπτωση αποτελεσμάτων ήταν αυτή της απλοτυπικής ανάλυσης του γονιδίου του υποδοχέα της αδρεναλίνης (Zee, Cook et al. 2005). Εδώ παρατηρήσαμε στατιστική σημαντικότητα, αλλά υπήρξαν διαφορετικοί X^2 μεταξύ λογιστικής και πολυωνιμικής παλινδρόμησης (εικόνα 53). Κάτι τέτοιο είναι αναμενόμενο καθώς και σε αυτήν την περίπτωση υπάρχουν απλότυποι με μηδενική συχνότητα. Ωστόσο κανένας από τους απλότυπους δεν έδειξε σημαντική συσχέτιση και τα p-value ήταν όλα παραπάνω από την περιοχή απόρριψης. Θα μπορούσαμε να πούμε για τον απλότυπο 2 ότι έχει μια μικρή τάση καθώς είναι μόλις 0.0061 μονάδες εκτός της περιοχής απόρριψης. Ίσως αν είχαμε μεγαλύτερο δείγμα να μπορούσαμε να βγάλουμε ένα καλύτερο συμπέρασμα για αυτόν τον απλότυπο. Τα δεδομένα που επεξεργαστήκαμε φαίνονται στον πίνακα 13.

haplotype	haplotype_nr	frequency	case
G16-E27-T164	1	223	1
G16-Q27-I164	2	1	1
G16-Q27-T164	3	89	1
G16-E27-I164	4	0	1
R16-Q27-T164	5	209	1
R16-Q27-I164	6	0	1
R16-E27-T164	7	1	1
R16-E27-I164	8	0	1
G16-E27-T164	1	858	0
G16-Q27-I164	2	27	0
G16-Q27-T164	3	414	0
G16-E27-I164	4	0	0
R16-Q27-T164	5	791	0
R16-Q27-I164	6	0	0
R16-E27-T164	7	0	0
R16-E27-I164	8	0	0

Πίνακας 13: Απλοτυπικά δεδομένα ασθενών-μαρτύρων

Ακόμη μια οριακή περίπτωση αποτελεσμάτων είχαμε και σε ακόμη μια μελέτη η οποία αφορούσε εκτίμηση γενετικού κινδύνου για την διαβητική νεφροπάθεια (Kankova, Stejskalova et al. 2007). Αυτή τη φορά όμως είχαμε οριακό αριθμό p-value στον έλεγχο X^2. Είχαμε υψηλότερο X^2 από την προηγούμενη μελέτη (Zee, Cook et al. 2005), αλλά το p-value είναι πάνω από την περιοχή απόρριψης κατά 0.0004 μονάδες. Συνεπώς θα πρέπει να αναφερθούμε σε αυτήν την μελέτη. Βλέπουμε στην εικόνα 55 ότι δυο απλότυποι εμφανίζουν σημαντική συσχέτιση, οι 4 και 5 έχουν p-value<0.05. Επαναλαμβάνουμε την λογιστική παλινδρόμηση και εκτυπώνουμε τα αποτελέσματα με Odds ratio στην εικόνα 54.

Βλέπουμε λοιπόν πως οι απλότυποι 4 και 5 έχουν τον ίδιο σχετικό κίνδυνο για την ασθένεια, και αυτός είναι 2.21 φορές παραπάνω από τον απλότυπο με την μεγαλύτερη συχνότητα, δηλαδή τον πρώτο. Τα δεδομένα που επεξεργαστήκαμε φαίνονται παρακάτω στον πίνακα 14.

haplotype	haplotype_nr	frequency	case
1111	1	90	1
1211	2	43	1
2111	3	30	1
1122	4	25	1
2211	5	19	1
1222	6	12	1
2122	7	8	1
2222	8	4	1
rare	9	5	1
1111	1	199	0
1211	2	68	0
2111	3	52	0
1122	4	25	0
2211	5	19	0
1222	6	18	0
2122	7	7	0
2222	8	5	0
rare	9	19	0

Πίνακας 14: Απλοτυπικά δεδομένα ασθενών-μαρτύρων

```
Logistic regression                                     Number of obs   =       2612
                                                        LR chi2(3)      =       9.13
                                                        Prob > chi2     =     0.0276
Log likelihood = -1301.9317                             Pseudo R2       =     0.0035

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -1.948405    1.02112    -1.91   0.056    -3.949764    .0529546
_Ihaplotyp~3 |  -.1897973   .1389288    -1.37   0.172    -.4620928    .0824982
_Ihaplotyp~5 |   .0164686   .1081606     0.15   0.879    -.1955223    .2284595
       _cons |  -1.347432   .0751652   -17.93   0.000    -1.494753   -1.200111
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       2613
                                                        LR chi2(4)      =      12.35
                                                        Prob > chi2     =     0.0149
Log likelihood = -1301.9317                             Pseudo R2       =     0.0047

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -1.948405    1.02112    -1.91   0.056    -3.949764    .0529546
_Ihaplotyp~3 |  -.1897973   .1389288    -1.37   0.172    -.4620928    .0824982
_Ihaplotyp~5 |   .0164686   .1081606     0.15   0.879    -.1955223    .2284595
_Ihaplotyp~7 |   36.00055   3.36e+07     0.00   1.000    -6.58e+07    6.58e+07
       _cons |  -1.347432   .0751652   -17.93   0.000    -1.494753   -1.200111
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         16
                                                        LR chi2(15)     =    6129.17
                                                        Prob > chi2     =     0.0000
Log likelihood = -27.431643                             Pseudo R2       =     0.9911

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -1.347441   .0751651   -17.93   0.000    -1.494762    -1.20012
_Ihaplotyp~2 |  -3.458776   .1954545   -17.70   0.000     -3.84186   -3.075692
_Ihaplotyp~3 |  -.7287422   .0598409   -12.18   0.000    -.8460282   -.6114562
_Ihaplotyp~4 |  -23.61996   4594.774    -0.01   0.996    -9029.212    8981.972
_Ihaplotyp~5 |  -.0813149   .0492921    -1.65   0.099    -.1779256    .0152958
_Ihaplotyp~6 |  -23.61996   4594.774    -0.01   0.996    -9029.212    8981.972
_Ihaplotyp~7 |  -23.61996   4594.774    -0.01   0.996    -9029.212    8981.972
_Ihaplotyp~8 |  -23.61996   4594.774    -0.01   0.996    -9029.212    8981.972
_IcasXhap_~2 |  -1.948395   1.021119    -1.91   0.056    -3.949752    .0529618
_IcasXhap_~3 |  -.1897856   .1389283    -1.37   0.172    -.4620801     .082509
_IcasXhap_~4 |   1.347441   6497.992     0.00   1.000    -12734.48    12737.18
_IcasXhap_~5 |   .0164764   .1081605     0.15   0.879    -.1955142    .2284671
_IcasXhap_~6 |   1.347441   6497.992     0.00   1.000    -12734.48    12737.18
_IcasXhap_~7 |   18.21279   4594.774     0.00   0.997    -8987.379    9023.805
_IcasXhap_~8 |   1.347441   6497.992     0.00   1.000    -12734.48    12737.18
       _cons |   6.754615   .0341393   197.85   0.000     6.687703    6.821526
------------------------------------------------------------------------------
```

Εικόνα 53: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Haplotype Analysis of the β2 Adrenergic Receptor Gene and Risk of Myocardial Infarction in Humans" (Zee, Cook et al. 2005)

```
Logistic regression                                     Number of obs   =        648
                                                        LR chi2(8)      =      15.49
                                                        Prob > chi2     =     0.0504
Log likelihood = -417.21246                             Pseudo R2       =     0.0182

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   1.398203   .3252074     1.44   0.150     .8863178    2.205722
_Ihaplotyp~3 |   1.275641   .3343551     0.93   0.353     .7631716    2.132234
_Ihaplotyp~4 |   2.211111   .6855737     2.56   0.010     1.204171    4.060063
_Ihaplotyp~5 |   2.211111   .7704045     2.28   0.023     1.116944    4.377133
_Ihaplotyp~6 |   1.474074   .5803906     0.99   0.324     .6813555    3.189076
_Ihaplotyp~7 |   2.526984   1.346656     1.74   0.082     .8891775    7.181523
_Ihaplotyp~8 |   1.768889   1.207694     0.84   0.404      .464034    6.742972
_Ihaplotyp~9 |   .5818713   .3016583    -1.04   0.296     .2106398    1.607361
------------------------------------------------------------------------------
```

Εικόνα 54: Απεικόνιση των συντελεστών με Odds ratio – επέκταση της εικόνας 54

```
Logistic regression                                     Number of obs   =        648
                                                        LR chi2(8)      =      15.49
                                                        Prob > chi2     =     0.0504
Log likelihood = -417.21246                             Pseudo R2       =     0.0182

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .3351876   .2325896     1.44   0.150    -.1206797    .7910549
_Ihaplotyp~3 |   .2434488   .2621075     0.93   0.353    -.2702724    .7571701
_Ihaplotyp~4 |   .7934952   .3100584     2.56   0.010     .1857918    1.401199
_Ihaplotyp~5 |   .7934952   .3484242     2.28   0.023     .1105964    1.476394
_Ihaplotyp~6 |     .38803   .3937323     0.99   0.324    -.3836711    1.159731
_Ihaplotyp~7 |   .9270265   .5329103     1.74   0.082    -.1174584    1.971512
_Ihaplotyp~8 |   .5703516   .6827417     0.84   0.404    -.7677976    1.908501
_Ihaplotyp~9 |  -.5415059   .5184278    -1.04   0.296    -1.557606     .474594
       _cons |  -.7934952   .1270285    -6.25   0.000    -1.042466   -.5445239
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        648
                                                        LR chi2(8)      =      15.49
                                                        Prob > chi2     =     0.0504
Log likelihood = -417.21246                             Pseudo R2       =     0.0182

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .3351876   .2325896     1.44   0.150    -.1206797    .7910549
_Ihaplotyp~3 |   .2434488   .2621075     0.93   0.353    -.2702724    .7571701
_Ihaplotyp~4 |   .7934952   .3100584     2.56   0.010     .1857918    1.401199
_Ihaplotyp~5 |   .7934952   .3484242     2.28   0.023     .1105964    1.476394
_Ihaplotyp~6 |     .38803   .3937323     0.99   0.324    -.3836711    1.159731
_Ihaplotyp~7 |   .9270265   .5329103     1.74   0.082    -.1174584    1.971512
_Ihaplotyp~8 |   .5703516   .6827417     0.84   0.404    -.7677976    1.908501
_Ihaplotyp~9 |  -.5415059   .5184278    -1.04   0.296    -1.557606     .474594
       _cons |  -.7934952   .1270285    -6.25   0.000    -1.042466   -.5445239
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         18
                                                        LR chi2(17)     =     709.82
                                                        Prob > chi2     =     0.0000
Log likelihood = -43.859131                             Pseudo R2       =     0.8900

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.7934952   .1270285    -6.25   0.000    -1.042466   -.5445239
_Ihaplotyp~2 |  -1.073797   .1404671    -7.64   0.000    -1.349108   -.7984866
_Ihaplotyp~3 |  -1.342061    .155743    -8.62   0.000    -1.647312    -1.03681
_Ihaplotyp~4 |  -2.074429   .2121912    -9.78   0.000    -2.490316   -1.658542
_Ihaplotyp~5 |  -2.348866   .2401181    -9.78   0.000    -2.819489   -1.878243
_Ihaplotyp~6 |  -2.402933   .2461314    -9.76   0.000    -2.885342   -1.920524
_Ihaplotyp~7 |  -3.347395   .3845546    -8.70   0.000    -4.101108   -2.593681
_Ihaplotyp~8 |  -3.683867    .452797    -8.14   0.000    -4.571333   -2.796401
_Ihaplotyp~9 |  -2.348866   .2401181    -9.78   0.000    -2.819489   -1.878243
_IcasXhap_~2 |   .3351876   .2325896     1.44   0.150    -.1206797    .7910549
_IcasXhap_~3 |   .2434488   .2621075     0.93   0.353    -.2702724    .7571701
_IcasXhap_~4 |   .7934952   .3100584     2.56   0.010     .1857918    1.401199
_IcasXhap_~5 |   .7934952   .3484242     2.28   0.023     .1105964    1.476394
_IcasXhap_~6 |     .38803   .3937323     0.99   0.324    -.3836711    1.159731
_IcasXhap_~7 |   .9270265   .5329103     1.74   0.082    -.1174584    1.971512
_IcasXhap_~8 |   .5703516   .6827417     0.84   0.404    -.7677976    1.908501
_IcasXhap_~9 |  -.5415059   .5184282    -1.04   0.296    -1.557607    .4745947
       _cons |   5.293305   .0708881    74.67   0.000     5.154367    5.432243
------------------------------------------------------------------------------
```

Εικόνα 55: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Genetic risk factors for diabetic nephropathy on chromosomes 6p and 7q identified by the set-association approach" (Kankova, Stejskalova et al. 2007)

3.2 ΣΥΖΗΤΗΣΗ

Τα αποτελέσματα αυτής της εργασίας συμφώνησαν με αυτά των αντίστοιχων μελετών. Ο αριθμός των μελετών που κατέληγαν σε κάποιο τρανταχτό συμπέρασμα ήταν μικρός, αυτό οφείλεται στο ότι οι μελέτες αυτού του τύπου είναι κυρίως ερευνητικές και έτσι είναι νωρίς να γίνεται λόγος για σίγουρα αποτελέσματα, εκτός και αν τα στοιχεία είναι πολύ ενθαρρυντικά. Στα σημεία που υπήρξαν στατιστικώς σημαντικά αποτελέσματα, ήρθαν σε συμφωνία με τα δικά μας, και αυτό απέδειξε πως η μέθοδος της παλινδρόμησης αποτέλεσε ένα καλό και αξιόπιστο εργαλείο για την ανάλυση των δεδομένων.

Ένα βασικό συμπέρασμα είναι ότι πρέπει κανείς να προσέχει την οργάνωση των δεδομένων στους πίνακες. Η κωδικοποίηση που γίνεται στο Stata έχει την αρχή να δεσμεύει τον πρώτο απλότυπο και να τον χρησιμοποιεί ως αυτόν με τη μεγαλύτερη συχνότητα. Σε αρκετές μελέτες όμως τα δεδομένα δεν ήταν ταξινομημένα κατά αύξουσα συχνότητα, και ως συνέπεια έπρεπε να αλλάξουν κατά την μεταφορά. Επίσης προσοχή χρειάζεται στην ομοιογένεια των δεδομένων. Ο τρόπος απεικόνισης μπορεί να διαφέρει από μελέτη σε μελέτη, καθώς και οι μονάδες μέτρησης.

Σε πολλές περιπτώσεις χρειάστηκαν αλλαγές κυρίως στην μετατροπή των συχνοτήτων, αλλά και τις τιμές της χοληστερίνης στο αίμα. Για την ανάλυση των μελετών με μεταβλητή συνεχούς τιμής είχαμε σαν κατάλληλο εργαλείο την απλή γραμμική παλινδρόμηση, ωστόσο τα δεδομένα ήταν λίγα και αυτή η μέθοδος δεν παρουσίαζε αποτελέσματα. Έτσι χρησιμοποιήσαμε έναν άλλο τύπο παλινδρόμησης που παρέχει το Stata, την λεγόμενη vwls (variance weighted least squares) παλινδρόμηση, η οποία ταίριαξε στα δεδομένα καθώς θέλαμε να χρησιμοποιήσουμε ως παράγοντα και την διακύμανση των τιμών της χοληστερίνης.

Όσον αφορά τις μελέτες με ασθενείς και μάρτυρες τα αποτελέσματα ήταν ενθαρρυντικά και έδειξαν πως με αυτές τις απλές μεθόδους μπορούμε να κάνουμε ανάλυση απλότυπων, αλλά και παρόμοιων δεδομένων όπως είναι οι γονότυποι αλλά και μεμονωμένοι πολυμορφισμοί. Η πιο κατάλληλη για την παρουσίαση των αποτελεσμάτων αποδείχθηκε η πολυωνιμική παλινδρόμηση λόγω της δομής της, ωστόσο η διαφορά της με την λογιστική δεν είναι μεγάλη οπτικά. Από την υπολογιστική όμως πλευρά η πολυωνυμική παλινδρόμηση έδειξε ένα βασικό πλεονέκτημα, το οποίο ήταν ο υπολογισμός των συντελεστών παλινδρόμησης με μεγαλύτερη ακρίβεια. Σε δυο περιπτώσεις η λογιστική παλινδρόμηση εξαίρεσε

κάποιους απλότυπους από την διαδικασία επειδή είχαν μηδενική συχνότητα, ενώ η πολυωνυμική όχι. Η Poisson αρκετές φορές δεν συμφωνούσε στους υπολογισμούς του p-value για τον X^2, αλλά ούτε για τον ίδιο. Ωστόσο είχε το ίδιο αποτέλεσμα με τις υπόλοιπες στους συντελεστές παλινδρόμησης και στα αντίστοιχα p-value. Αυτό ίσως μας παροτρύνει να την χρησιμοποιούμε συνδιαστικά με τις άλλες, αλλά ίσως και καθόλου, ανάλογα κάθε περίπτωση.

Πολύ σημαντικός παράγοντας που επηρεάζει την αξιοπιστία των αποτελεσμάτων είναι το μέγεθος του δείγματος. Ένας ικανοποιητικός αριθμός ατόμων μπορεί να συμβάλει στην εξαγωγή αξιόπιστων αποτελεσμάτων. Στις μελέτες που είδαμε ο μέσος αριθμός των ατόμων ήταν 1000-1200. Υπήρξαν βέβαια περιπτώσεις πολύ κάτω από αυτό το πλαίσιο, αλλά και πολύ παραπάνω. Σε όσες περιπτώσεις είχαμε πολλά παραπάνω άτομα, παρατηρήσαμε πως αποφεύγαμε τις οριακές καταστάσεις των p-value και το αποτέλεσμα ήταν πιο ξεκάθαρο.

Αυτές οι παλινδρομήσεις είναι η βάση στην οποία μπορεί να στηριχθεί κανείς στην ανάλυση τέτοιων δεδομένων, και αποτελούν εξέλιξη του υπολογισμού του σχετικού λόγου πιθανοτήτων (Odds ratio). Για την ακρίβεια στηρίζονται πάνω στο μαθηματικό αυτό εργαλείο. Αν λάβουμε υπόψη τα πρόσθετα που μπορούμε να βάλουμε σε αυτές τις μεθόδους, όπως τα πιθανοτικά βάρη, θα έχουμε στα χέρια μας απλά και χρήσιμα εργαλεία για απλοτυπική ανάλυση.

Επιπλέον, σημαντικό είναι να ξέρουμε πως δεν μπορούμε να βγάλουμε εύκολα συμπέρασμα για μια ασθένεια, εάν οφείλεται ή όχι σε έναν απλότυπο, πολύ απλά γιατί υπάρχουν και άλλοι παράγοντες που επηρεάζουν την πορεία της. Οι απλότυποι αποτελούν ένα τμήμα της ανάλυσης του ανθρώπινου γονιδιώματος. Για την εξαγωγή κάποιου συμπεράσματος χρειάζεται αρκετός χρόνος έρευνας και στατιστικών μελετών. Το συμπέρασμα αυτής της εργασίας είναι πως θα πρέπει να συμπεριλαμβάνουμε στην μελέτη γενετικής συσχέτισης την ανάλυση των απλότυπων ως ένα παραπάνω στοιχείο που θα οδηγήσει σε κάποιο συμπέρασμα. Η βελτιστοποίηση αυτής της ανάλυσης ελπίζουμε πως θα οδηγήσει σε πιο αξιόπιστα και αληθή αποτελέσματα στο μέλλον.

Η σημαντικότητα της απλοτυπικής ανάλυσης φαίνεται καθώς πολλές μελέτες πλέον την συμπεριλαμβάνουν στην δουλειά τους. Από την πρώτη εμφάνιση των απλοτυπικών αναλύσεων την δεκαετία του ’90 και μέχρι σήμερα παρατηρούμε έναν αυξανόμενο αριθμό μελετών που περιέχουν τέτοια δεδομένα. Το μέλλον είναι αρκετά

υποσχόμενο καθώς ήδη αποτυπικά δεδομένα είναι διαθέσιμα και στο διαδίκτυο μέσω του HapMap project, πράγμα που ενθαρρύνει την ενασχόληση φοιτητών/ερευνητών με το θέμα.

Ακολουθούν όλες οι υπόλοιπες μελέτες που δεν αναλύθηκαν. Τις χωρίσαμε σε 2 κατηγορίες. Η πρώτη κατηγορία είναι μελέτες με απόρριψη της μηδενικής υπόθεσης (p-value<0.05) του X^2 και η δεύτερη με αποδοχή της μηδενικής υπόθεσης (p-value>0.05) του X^2.

Παράρτημα 1 – απόρριψη της μηδενικής υπόθεσης

haplotype	haplotype_nr	frequency	case
A 274 T C	1	272	1
A 274 C T	2	19	1
G 270 C T	3	175	1
G 270 T C	4	25	1
G 274 C T	5	77	1
A 274 T C	1	285	0
A 274 C T	2	47	0
G 270 C T	3	163	0
G 270 T C	4	29	0
G 274 C T	5	75	0

Πίνακας 15: Δεδομένα της μελέτης "Brain-derived neurotrophic factor and risk of schizophrenia: An association study and meta-analysis" (Qian, Zhao et al. 2007)

```
Logistic regression                                     Number of obs   =       1167
                                                        LR chi2(4)      =      12.49
                                                        Prob > chi2     =     0.0140
Log likelihood = -802.24476                             Pseudo R2       =     0.0077

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.8590215   .2847691    -3.02   0.003    -1.417159   -.3008844
_Ihaplotyp~3 |   .1177229   .1379656     0.85   0.394    -.1526847    .3881304
_Ihaplotyp~4 |  -.1017329   .2857761    -0.36   0.722    -.6618438     .458378
_Ihaplotyp~5 |   .0730044   .1830453     0.40   0.690    -.2857578    .4317667
       _cons |  -.0466871   .0847658    -0.55   0.582     -.212825    .1194508
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1167
                                                        LR chi2(4)      =      12.49
                                                        Prob > chi2     =     0.0140
Log likelihood = -802.24476                             Pseudo R2       =     0.0077

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.8590215   .2847691    -3.02   0.003    -1.417159   -.3008844
_Ihaplotyp~3 |   .1177229   .1379656     0.85   0.394    -.1526847    .3881304
_Ihaplotyp~4 |  -.1017329   .2857761    -0.36   0.722    -.6618438     .458378
_Ihaplotyp~5 |   .0730044   .1830453     0.40   0.690    -.2857578    .4317667
       _cons |  -.0466871   .0847658    -0.55   0.582     -.212825    .1194508
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         10
                                                        LR chi2(9)      =     777.41
                                                        Prob > chi2     =     0.0000
Log likelihood =  -30.98497                             Pseudo R2       =     0.9262

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.0466871   .0847658    -0.55   0.582     -.212825    .1194508
_Ihaplotyp~2 |  -1.802342   .1574337   -11.45   0.000    -2.110906   -1.493777
_Ihaplotyp~3 |   -.558739   .0982026    -5.69   0.000    -.7512124   -.3662655
_Ihaplotyp~4 |  -2.285193   .1949142   -11.72   0.000    -2.667218   -1.903169
_Ihaplotyp~5 |  -1.335001   .1297771   -10.29   0.000     -1.58936   -1.080643
_IcasXhap_~2 |  -.8590215   .2847691    -3.02   0.003    -1.417159   -.3008844
_IcasXhap_~3 |   .1177229   .1379656     0.85   0.394    -.1526847    .3881304
_IcasXhap_~4 |  -.1017329   .2857761    -0.36   0.722    -.6618438     .458378
_IcasXhap_~5 |   .0730044   .1830453     0.40   0.690    -.2857578    .4317667
       _cons |   5.652489   .0592349    95.43   0.000     5.536391    5.768587
------------------------------------------------------------------------------
```

Εικόνα 56: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Brain-derived neurotrophic factor and risk of schizophrenia: An association study and meta-analysis" (Qian, Zhao et al. 2007)

haplotype	haplotype_nr	frequency	percent	case
GAGCCCA	1	7094	0,31	1
GAGCCCG	2	5950	0,26	1
GGCACCG	3	4577	0,2	1
GGCACTG	4	2060	0,09	1
TGCACCG	5	1831	0,08	1
GAGCCCA	1	3880	0,3	1
GAGCCCG	2	3363	0,26	1
GGCACCG	3	2587	0,2	1
GGCACTG	4	1035	0,08	1
TGCACCG	5	1164	0,09	1
GAGCCCA	1	9720	0,32	0
GAGCCCG	2	7594	0,25	0
GGCACCG	3	6379	0,21	0
GGCACTG	4	3037	0,1	0
TGCACCG	5	2430	0,08	0
GAGCCCA	1	3768	0,28	0
GAGCCCG	2	3499	0,26	0
GGCACCG	3	2692	0,2	0
GGCACTG	4	1077	0,08	0
TGCACCG	5	1211	0,09	0

Πίνακας 16: Δεδομένα της μελέτης "Genetic variation in IL6 gene and type 2 diabetes: tagging-SNP haplotype analysis in large-scale case–control study and meta-analysis" (Qi, van Dam et al. 2006)

```
Logistic regression                                     Number of obs   =      74948
                                                        LR chi2(4)      =      19.23
                                                        Prob > chi2     =     0.0007
Log likelihood = -51526.837                             Pseudo R2       =     0.0002

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .0313686    .019047     1.65   0.100    -.0059629       .0687
_Ihaplotyp~3 |  -.0297425   .0203738    -1.46   0.144    -.0696744    .0101895
_Ihaplotyp~4 |  -.0783363   .0270451    -2.90   0.004    -.1313438   -.0253289
_Ihaplotyp~5 |   .0109574   .0278173     0.39   0.694    -.0435635    .0654784
       _cons |  -.2062716   .0128555   -16.05   0.000    -.2314679   -.1810752
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =      74948
                                                        LR chi2(4)      =      19.23
                                                        Prob > chi2     =     0.0007
Log likelihood = -51526.837                             Pseudo R2       =     0.0002

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .0313686    .019047     1.65   0.100    -.0059629       .0687
_Ihaplotyp~3 |  -.0297425   .0203738    -1.46   0.144    -.0696744    .0101895
_Ihaplotyp~4 |  -.0783363   .0270451    -2.90   0.004    -.1313438   -.0253289
_Ihaplotyp~5 |   .0109574   .0278173     0.39   0.694    -.0435635    .0654784
       _cons |  -.2062716   .0128555   -16.05   0.000    -.2314679   -.1810752
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         20
                                                        LR chi2(9)      =   18619.43
                                                        Prob > chi2     =     0.0000
Log likelihood = -5067.7189                             Pseudo R2       =     0.6475

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Icase_1 |  -.2062716   .0128555   -16.05   0.000    -.2314679   -.1810752
_Ihaplotyp~2 |  -.1954861   .0128174   -15.25   0.000    -.2206079   -.1703644
_Ihaplotyp~3 |  -.3967179   .0135787   -29.22   0.000    -.4233317   -.3701041
_Ihaplotyp~4 |  -1.187405   .0178105   -66.67   0.000    -1.222312   -1.152497
_Ihaplotyp~5 |  -1.309542   .0186759   -70.12   0.000    -1.346146   -1.272938
_IcasXhap_~2 |   .0313686    .019047     1.65   0.100    -.0059629       .0687
_IcasXhap_~3 |  -.0297425   .0203738    -1.46   0.144    -.0696744    .0101895
_IcasXhap_~4 |  -.0783363   .0270451    -2.90   0.004    -.1313438   -.0253288
_IcasXhap_~5 |   .0109574   .0278173     0.39   0.694    -.0435635    .0654784
       _cons |   8.816408   .0086105  1023.92   0.000     8.799532    8.833285
------------------------------------------------------------------------------
```

Εικόνα 57: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Genetic variation in IL6 gene and type 2 diabetes: tagging-SNP haplotype analysis in large-scale case–control study and meta-analysis" (Qi, van Dam et al. 2006)

haplotype	haplotype_nr	frequency	case
CGGAT	1	355	1
CGGTT	2	163	1
CAGTT	3	45	1
AAGTT	4	35	1
AAATC	5	129	1
Others	6	45	1
CGGAT	1	286	0
CGGTT	2	185	0
CAGTT	3	63	0
AAGTT	4	21	0
AAATC	5	68	0
Others	6	129	0

Πίνακας 17: Δεδομένα της μελέτης "Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma" (Liu, Zhang et al. 2007) – για το γονίδιο XRCC5

```
Logistic regression                                     Number of obs   =       1524
                                                        LR chi2(5)      =      76.62
                                                        Prob > chi2     =     0.0000
Log likelihood =  -1017.915                             Pseudo R2       =     0.0363

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.3427316    .133618    -2.57   0.010    -.6046181   -.0808451
_Ihaplotyp~3 |  -.5525982   .2107336    -2.62   0.009    -.9656284    -.139568
_Ihaplotyp~4 |   .2946996   .2872349     1.03   0.305    -.2682704    .8576697
_Ihaplotyp~5 |   .4241787   .1696208     2.50   0.012     .091728    .7566294
_Ihaplotyp~6 |  -1.269276   .1904929    -6.66   0.000    -1.642635   -.8959166
       _cons |    .216126   .0794569     2.72   0.007     .0603932    .3718587
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1524
                                                        LR chi2(5)      =      76.62
                                                        Prob > chi2     =     0.0000
Log likelihood =  -1017.915                             Pseudo R2       =     0.0363

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.3427316    .133618    -2.57   0.010    -.6046181   -.0808451
_Ihaplotyp~3 |  -.5525982   .2107336    -2.62   0.009    -.9656284    -.139568
_Ihaplotyp~4 |   .2946996   .2872349     1.03   0.305    -.2682704    .8576697
_Ihaplotyp~5 |   .4241787   .1696208     2.50   0.012     .091728    .7566294
_Ihaplotyp~6 |  -1.269276   .1904929    -6.66   0.000    -1.642635   -.8959166
       _cons |    .216126   .0794569     2.72   0.007     .0603932    .3718587
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         12
                                                        LR chi2(11)     =     896.94
                                                        Prob > chi2     =     0.0000
Log likelihood = -38.111482                             Pseudo R2       =     0.9217

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |    .216126   .0794569     2.72   0.007     .0603932    .3718587
_Ihaplotyp~2 |   -.435636   .0943499    -4.62   0.000    -.6205584   -.2507135
_Ihaplotyp~3 |  -1.512857   .1391744   -10.87   0.000    -1.785634    -1.24008
_Ihaplotyp~4 |  -2.611469   .2260875   -11.55   0.000    -3.054593   -2.168346
_Ihaplotyp~5 |  -1.436484   .1349162   -10.65   0.000    -1.700915   -1.172053
_Ihaplotyp~6 |  -.7961794   .1060587    -7.51   0.000    -1.004051   -.5883082
_IcasXhap_~2 |  -.3427316    .133618    -2.57   0.010    -.6046181   -.0808451
_IcasXhap_~3 |  -.5525982   .2107336    -2.62   0.009    -.9656284    -.139568
_IcasXhap_~4 |   .2946996   .2872349     1.03   0.305    -.2682704    .8576697
_IcasXhap_~5 |   .4241787   .1696208     2.50   0.012     .091728    .7566294
_IcasXhap_~6 |  -1.269276    .190493    -6.66   0.000    -1.642635   -.8959166
       _cons |   5.655992   .0591312    95.65   0.000     5.540097    5.771887
------------------------------------------------------------------------------
```

Εικόνα 58: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma" (Liu, Zhang et al. 2007) – για το γονίδιο XRCC5

haplotype	haplotype_nr	frequency	case
CAGGG	1	348	1
CAAGG	2	198	1
CGAAA	3	56	1
GAAGG	4	130	1
Others	5	39	1
CAGGG	1	326	0
CAAGG	2	205	0
CGAAA	3	23	0
GAAGG	4	126	0
Others	5	71	0

Πίνακας 18: Δεδομένα της μελέτης "Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma"(Liu, Zhang et al. 2007) – για το γονίδιο XRCC6

```
Logistic regression                                     Number of obs   =       1522
                                                        LR chi2(4)      =      24.30
                                                        Prob > chi2     =     0.0001
Log likelihood =  -1042.688                             Pseudo R2       =     0.0115

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   -.100048   .1259746    -0.79   0.427    -.3469537    .1468576
_Ihaplotyp~3 |   .8245524    .259377     3.18   0.001     .3161827    1.332922
_Ihaplotyp~4 |  -.0340526   .1468668    -0.23   0.817    -.3219062    .2538011
_Ihaplotyp~5 |  -.6644233   .2136974    -3.11   0.002    -1.083263   -.2455841
       _cons |   .0653051   .0770782     0.85   0.397    -.0857654    .2163756
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1522
                                                        LR chi2(4)      =      24.30
                                                        Prob > chi2     =     0.0001
Log likelihood =  -1042.688                             Pseudo R2       =     0.0115

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   -.100048   .1259746    -0.79   0.427    -.3469537    .1468576
_Ihaplotyp~3 |   .8245524    .259377     3.18   0.001     .3161827    1.332922
_Ihaplotyp~4 |  -.0340526   .1468668    -0.23   0.817    -.3219062    .2538011
_Ihaplotyp~5 |  -.6644233   .2136974    -3.11   0.002    -1.083263   -.2455841
       _cons |   .0653051   .0770782     0.85   0.397    -.0857654    .2163756
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         10
                                                        LR chi2(9)      =     796.52
                                                        Prob > chi2     =     0.0000
Log likelihood = -32.721072                             Pseudo R2       =     0.9241

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .0653051   .0770782     0.85   0.397    -.0857654    .2163756
_Ihaplotyp~2 |  -.4638874   .0891377    -5.20   0.000    -.6385941   -.2891807
_Ihaplotyp~3 |  -2.651403   .2157446   -12.29   0.000    -3.074255   -2.228551
_Ihaplotyp~4 |  -.9506155   .1048999    -9.06   0.000    -1.156216   -.7450154
_Ihaplotyp~5 |  -1.524218   .1309656   -11.64   0.000    -1.780905    -1.26753
_IcasXhap_~2 |   -.100048   .1259746    -0.79   0.427    -.3469537    .1468576
_IcasXhap_~3 |   .8245524    .259377     3.18   0.001     .3161827    1.332922
_IcasXhap_~4 |  -.0340526   .1468668    -0.23   0.817    -.3219062    .2538011
_IcasXhap_~5 |  -.6644233   .2136974    -3.11   0.002    -1.083263   -.2455841
       _cons |   5.786897   .0553849   104.49   0.000     5.678345     5.89545
------------------------------------------------------------------------------
```

Εικόνα 59: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma"(Liu, Zhang et al. 2007) – για το γονίδιο XRCC6

haplotype	haplotype_nr	frequency	case
GCCT	1	550	1
AGTC	2	266	1
GCTC	3	31	1
ACTC	4	40	1
ACCT	5	11	1
7 others	6	17	1
GCCT	1	558	0
AGTC	2	547	0
GCTC	3	35	0
ACTC	4	34	0
ACCT	5	19	0
7 others	6	23	0

Πίνακας 19: Δεδομένα της μελέτης "Haplotypic variation in MRE11, RAD50 and NBS1 and risk of non-Hodgkin's lymphoma" (Rollinson, Kesby et al. 2006)

```
Logistic regression                                     Number of obs   =       2131
                                                        LR chi2(5)      =      60.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -1425.5938                             Pseudo R2       =     0.0207

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.7065118   .0959048    -7.37   0.000    -.8944817   -.5185419
_Ihaplotyp~3 |  -.1069202   .2538499    -0.42   0.674    -.6044569    .3906166
_Ihaplotyp~4 |   .1769596   .2408777     0.73   0.463    -.2951519    .6490712
_Ihaplotyp~5 |   -.532103   .3836024    -1.39   0.165     -1.28395    .2197438
_Ihaplotyp~6 |  -.2878402   .3254414    -0.88   0.376    -.9256936    .3500132
       _cons |  -.0144407   .0600857    -0.24   0.810    -.1322066    .1033252
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       2131
                                                        LR chi2(5)      =      60.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -1425.5938                             Pseudo R2       =     0.0207

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.7065118   .0959048    -7.37   0.000    -.8944817   -.5185419
_Ihaplotyp~3 |  -.1069202   .2538499    -0.42   0.674    -.6044569    .3906166
_Ihaplotyp~4 |   .1769596   .2408777     0.73   0.463    -.2951519    .6490712
_Ihaplotyp~5 |   -.532103   .3836024    -1.39   0.165     -1.28395    .2197438
_Ihaplotyp~6 |  -.2878402   .3254414    -0.88   0.376    -.9256936    .3500132
       _cons |  -.0144407   .0600857    -0.24   0.810    -.1322066    .1033252
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         12
                                                        LR chi2(11)     =    3193.50
                                                        Prob > chi2     =     0.0000
Log likelihood = -36.077035                             Pseudo R2       =     0.9779

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.0144407   .0600857    -0.24   0.810    -.1322066    .1033252
_Ihaplotyp~2 |  -.0199102   .0601687    -0.33   0.741    -.1378386    .0980183
_Ihaplotyp~3 |  -2.769011   .1742514   -15.89   0.000    -3.110537   -2.427484
_Ihaplotyp~4 |  -2.797998   .1766462   -15.84   0.000    -3.144219   -2.451778
_Ihaplotyp~5 |   -3.37992   .2332889   -14.49   0.000    -3.837158   -2.922682
_Ihaplotyp~6 |  -3.188865   .2127684   -14.99   0.000    -3.605883   -2.771846
_IcasXhap_~2 |  -.7065118   .0959054    -7.37   0.000     -.894483   -.5185407
_IcasXhap_~3 |  -.1069202   .2538499    -0.42   0.674    -.6044569    .3906166
_IcasXhap_~4 |   .1769596   .2408777     0.73   0.463     -.295152    .6490712
_IcasXhap_~5 |   -.532103   .3836026    -1.39   0.165     -1.28395    .2197443
_IcasXhap_~6 |  -.2878402   .3254414    -0.88   0.376    -.9256936    .3500132
       _cons |   6.324359   .0423334   149.39   0.000     6.241387    6.407331
------------------------------------------------------------------------------
```

Εικόνα 60: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Haplotypic variation in MRE11, RAD50 and NBS1 and risk of non-Hodgkin's lymphoma" (Rollinson, Kesby et al. 2006)

haplotype	haplotype_nr	frequency	case
TTGGAG	1	135	1
CTGGGG	2	86	1
TAGGAG	3	81	1
TAGGAA	4	54	1
TTGTAG	5	25	1
TTSGAG	6	5	1
TTGGAA	7	4	1
CTGGAG	8	0	1
TTGGGG	9	3	1
Others	10	3	1
TTGGAG	1	185	0
CTGGGG	2	63	0
TAGGAG	3	109	0
TAGGAA	4	47	0
TTGTAG	5	21	0
TTSGAG	6	5	0
TTGGAA	7	2	0
CTGGAG	8	15	0
TTGGGG	9	6	0
Others	10	3	0

Πίνακας 20: Δεδομένα της μελέτης "Haplotype analysis of the RAGE gene: identification of a haplotype marker for diabetic nephropathy in type 2 diabetes mellitus" (Kankova, Stejskalova et al. 2005)

```
Logistic regression                                     Number of obs   =        837
                                                        LR chi2(8)      =      15.66
                                                        Prob > chi2     =     0.0475
Log likelihood = -571.12251                             Pseudo R2       =     0.0135

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .6262936   .2007828     3.12   0.002     .2327665    1.019821
_Ihaplotyp~3 |   .0181823   .1852911     0.10   0.922    -.3449816    .3813463
_Ihaplotyp~4 |   .4539175   .2293642     1.98   0.048      .004372    .903463
_Ihaplotyp~5 |   .4894344   .3169099     1.54   0.122    -.1316975    1.110566
_Ihaplotyp~6 |    .315081   .6425051     0.49   0.624    -.9442058    1.574368
_Ihaplotyp~7 |   1.008228   .8733916     1.15   0.248    -.7035878    2.720044
_Ihaplotyp~9 |  -.3780661   .7161095    -0.53   0.598    -1.781615    1.025483
_Ihaploty~10 |    .315081   .8243055     0.38   0.702    -1.300528     1.93069
       _cons |   -.315081   .1131937    -2.78   0.005    -.5369366   -.0932255
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        852
                                                        LR chi2(9)      =      34.65
                                                        Prob > chi2     =     0.0001
Log likelihood = -571.12251                             Pseudo R2       =     0.0294

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .6262936   .2007828     3.12   0.002     .2327665    1.019821
_Ihaplotyp~3 |   .0181823   .1852911     0.10   0.922    -.3449816    .3813463
_Ihaplotyp~4 |   .4539175   .2293642     1.98   0.048      .004372    .903463
_Ihaplotyp~5 |   .4894344   .3169099     1.54   0.122    -.1316975    1.110566
_Ihaplotyp~6 |    .315081   .6425051     0.49   0.624    -.9442058    1.574368
_Ihaplotyp~7 |   1.008228   .8733916     1.15   0.248    -.7035878    2.720044
_Ihaplotyp~8 |  -32.89549    4202796    -0.00   1.000     -8237362     8237297
_Ihaplotyp~9 |  -.3780661   .7161095    -0.53   0.598    -1.781615    1.025483
_Ihaploty~10 |    .315081   .8243055     0.38   0.702    -1.300528     1.93069
       _cons |   -.315081   .1131937    -2.78   0.005    -.5369366   -.0932255
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         20
                                                        LR chi2(19)     =    1135.70
                                                        Prob > chi2     =     0.0000
Log likelihood =  -45.17619                             Pseudo R2       =     0.9263

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.3151144   .1131929    -2.78   0.005    -.5369684   -.0932604
_Ihaplotyp~2 |  -1.077249   .1458703    -7.38   0.000    -1.363149   -.7913482
_Ihaplotyp~3 |  -.5290242   .1207451    -4.38   0.000    -.7656803   -.2923682
_Ihaplotyp~4 |  -1.370255   .1633466    -8.39   0.000    -1.690409   -1.050102
_Ihaplotyp~5 |   -2.17584   .2302673    -9.45   0.000    -2.627156   -1.724525
_Ihaplotyp~6 |  -3.610947   .4532156    -7.97   0.000    -4.499233   -2.722661
_Ihaplotyp~7 |  -4.527211   .7109076    -6.37   0.000    -5.920564   -3.133858
_Ihaplotyp~8 |  -2.512306    .268458    -9.36   0.000    -3.038474   -1.986138
_Ihaplotyp~9 |  -3.428629   .4148154    -8.27   0.000    -4.241653   -2.615606
_Ihaploty~10 |  -4.121749   .5820045    -7.08   0.000    -5.262457   -2.981041
_IcasXhap_~2 |   .6263554   .2007811     3.12   0.002     .2328316    1.019879
_IcasXhap_~3 |   .0182305   .1852892     0.10   0.922    -.3449296    .3813905
_IcasXhap_~4 |   .4539568   .2293647     1.98   0.048     .0044101    .9035034
_IcasXhap_~5 |   .4894542   .3169068     1.54   0.122    -.1316717     1.11058
_IcasXhap_~6 |   .3151144   .6425037     0.49   0.624    -.9441697    1.574398
_IcasXhap_~7 |   1.008244   .8733806     1.15   0.248    -.7035507    2.720038
_IcasXhap_~8 |  -16.92857   1433.396    -0.01   0.991    -2826.333    2792.476
_IcasXhap_~9 |   -.378005   .7161028    -0.53   0.598    -1.781541    1.025531
_IcasXhap~10 |   .3151144    .824294     0.38   0.702    -1.300472    1.930701
       _cons |   5.220389   .0735202    71.01   0.000     5.076292    5.364486
------------------------------------------------------------------------------
```

Εικόνα 61: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Haplotype analysis of the RAGE gene: identification of a haplotype marker for diabetic nephropathy in type 2 diabetes mellitus" (Kankova, Stejskalova et al. 2005)

Παράρτημα 2 – αποδοχή της μηδενικής υπόθεσης

haplotype	haplotype_nr	case	frequency
GTG	1	1	177
TCGA	2	1	128
CCCC	3	1	26
TCGC	4	1	5
CGCC	5	1	167
AGA	6	1	126
GGG	7	1	23
GTA	8	1	9
GTG	1	0	256
TCGA	2	0	168
CCCC	3	0	36
TCGC	4	0	9
CGCC	5	0	219
AGA	6	0	156
GGG	7	0	23
GTA	8	0	11

Πίνακας 21: Δεδομένα της μελέτης "Analysis of Candidate Genes on Chromosomes 5q and 19p in Celiac Disease" (Latiano, Mora et al. 2007)

```
Logistic regression                                     Number of obs   =       1539
                                                        LR chi2(7)      =       2.42
                                                        Prob > chi2     =     0.9330
Log likelihood = -1050.1936                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |    .097094   .1527117     0.64   0.525    -.2022153    .3964033
_Ihaplotyp~3 |   .0436053   .2753094     0.16   0.874    -.4959913    .5832019
_Ihaplotyp~4 |   -.218759   .5662739    -0.39   0.699    -1.328635    .8911175
_Ihaplotyp~5 |   .0979498   .1418104     0.69   0.490    -.1799936    .3758932
_Ihaplotyp~6 |   .1554536   .1546051     1.01   0.315    -.1475668     .458474
_Ihaplotyp~7 |   .3690277   .3106646     1.19   0.235    -.2398637    .9779191
_Ihaplotyp~8 |    .168357   .4599741     0.37   0.714    -.7331756     1.06989
       _cons |  -.3690277   .0977546    -3.78   0.000    -.5606233   -.1774322
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1539
                                                        LR chi2(7)      =       2.42
                                                        Prob > chi2     =     0.9330
Log likelihood = -1050.1936                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |    .097094   .1527117     0.64   0.525    -.2022153    .3964033
_Ihaplotyp~3 |   .0436053   .2753094     0.16   0.874    -.4959913    .5832019
_Ihaplotyp~4 |   -.218759   .5662739    -0.39   0.699    -1.328635    .8911175
_Ihaplotyp~5 |   .0979498   .1418104     0.69   0.490    -.1799936    .3758932
_Ihaplotyp~6 |   .1554536   .1546051     1.01   0.315    -.1475668     .458474
_Ihaplotyp~7 |   .3690277   .3106646     1.19   0.235    -.2398637    .9779191
_Ihaplotyp~8 |    .168357   .4599741     0.37   0.714    -.7331756     1.06989
       _cons |  -.3690277   .0977546    -3.78   0.000    -.5606233   -.1774322
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         16
                                                        LR chi2(15)     =    1308.16
                                                        Prob > chi2     =     0.0000
Log likelihood = -46.064986                             Pseudo R2       =     0.9342

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.3690277   .0977546    -3.78   0.000    -.5606233   -.1774322
_Ihaplotyp~2 |  -.4212135   .0992906    -4.24   0.000    -.6158195   -.2266074
_Ihaplotyp~3 |  -1.961659   .1780001   -11.02   0.000    -2.310532   -1.612785
_Ihaplotyp~4 |  -3.347953   .3391421    -9.87   0.000    -4.012659   -2.683247
_Ihaplotyp~5 |  -.1561057    .092046    -1.70   0.090    -.3365125    .0243011
_Ihaplotyp~6 |  -.4953214   .1015702    -4.88   0.000    -.6943954   -.2962475
_Ihaplotyp~7 |  -2.409683   .2176798   -11.07   0.000    -2.836328   -1.983039
_Ihaplotyp~8 |  -3.147282    .307921   -10.22   0.000    -3.750796   -2.543768
_IcasXhap_~2 |    .097094   .1527117     0.64   0.525    -.2022153    .3964033
_IcasXhap_~3 |   .0436053   .2753094     0.16   0.874    -.4959913    .5832019
_IcasXhap_~4 |  -.218759   .5662747    -0.39   0.699    -1.328637    .8911191
_IcasXhap_~5 |   .0979498   .1418104     0.69   0.490    -.1799936    .3758932
_IcasXhap_~6 |   .1554536   .1546051     1.01   0.315    -.1475668     .458474
_IcasXhap_~7 |   .3690277   .3106646     1.19   0.235    -.2398637    .9779191
_IcasXhap_~8 |    .168357   .4599741     0.37   0.714    -.7331756     1.06989
       _cons |   5.545177     .0625    88.72   0.000     5.42268    5.667675
```

Εικόνα 62: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Analysis of Candidate Genes on Chromosomes 5q and 19p in Celiac Disease" (Latiano, Mora et al. 2007)

haplotype	haplotype_nr	Frequency	frequency_percent	case
GGGCA	1	1147	0,621	1
GGTTG	2	298	0,1613	1
GGGCG	3	163	0,0881	1
AGGCA	4	114	0,0675	1
GAGCA	5	105	0,0568	1
GGGCA	1	529	0,6347	0
GGTTG	2	132	0,1589	0
GGGCG	3	73	0,088	0
AGGCA	4	51	0,0615	0
GAGCA	5	44	0,0526	0

Πίνακας 22: Δεδομένα της μελέτης "Progesterone receptor polymorphisms and risk of breast cancer:
results from two Australian breast cancer studies" (Johnatty, Spurdle et al. 2008)

```
Logistic regression                                     Number of obs   =       2656
                                                        LR chi2(4)      =       0.36
                                                        Prob > chi2     =     0.9854
Log likelihood = -1648.6345                             Pseudo R2       =     0.0001

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .0403749   .1170199     0.35   0.730    -.1889799    .2697297
_Ihaplotyp~3 |   .0293741    .150319     0.20   0.845    -.2652457    .3239939
_Ihaplotyp~4 |   .0304561   .1764709     0.17   0.863    -.3154204    .3763327
_Ihaplotyp~5 |    .095854   .1871184     0.51   0.608    -.2708912    .4625993
       _cons |   .7739167   .0525566    14.73   0.000     .6709076    .8769258
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       2656
                                                        LR chi2(4)      =       0.36
                                                        Prob > chi2     =     0.9854
Log likelihood = -1648.6345                             Pseudo R2       =     0.0001

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |
_Ihaplotyp~2 |  -.0403749   .1170199    -0.35   0.730    -.2697297    .1889799
_Ihaplotyp~3 |  -.0293741    .150319    -0.20   0.845    -.3239939    .2652457
_Ihaplotyp~4 |  -.0304561   .1764709    -0.17   0.863    -.3763327    .3154204
_Ihaplotyp~5 |   -.095854   .1871184    -0.51   0.608    -.4625993    .2708912
       _cons |  -.7739167   .0525566   -14.73   0.000    -.8769258   -.6709076
------------------------------------------------------------------------------
(case==1 is the base outcome)

Poisson regression                                      Number of obs   =         10
                                                        LR chi2(9)      =    2906.92
                                                        Prob > chi2     =     0.0000
Log likelihood = -34.390318                             Pseudo R2       =     0.9769

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Icase_1 |   .7739167   .0525566    14.73   0.000     .6709076    .8769258
_Ihaplotyp~2 |  -1.388187    .097294   -14.27   0.000    -1.578879   -1.197494
_Ihaplotyp~3 |  -1.980529   .1248559   -15.86   0.000    -2.225242   -1.735816
_Ihaplotyp~4 |  -2.339163   .1466227   -15.95   0.000    -2.626538   -2.051788
_Ihaplotyp~5 |  -2.486799   .1569001   -15.85   0.000    -2.794317    -2.17928
_IcasXhap_~2 |   .0403749   .1170199     0.35   0.730    -.1889799    .2697297
_IcasXhap_~3 |   .0293741    .150319     0.20   0.845    -.2652457    .3239939
_IcasXhap_~4 |   .0304561   .1764709     0.17   0.863    -.3154204    .3763327
_IcasXhap_~5 |    .095854   .1871184     0.51   0.608    -.2708912    .4625993
       _cons |   6.270988   .0434783   144.23   0.000     6.185773    6.356204
------------------------------------------------------------------------------
```

Εικόνα 63: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Progesterone receptor polymorphisms and risk of breast cancer:
results from two Australian breast cancer studies" (Johnatty, Spurdle et al. 2008)

haplotype	haplotype_nr	frequency	case
CGC	1	28	1
TTT	2	24	1
CGT	3	2	1
TGC	4	6	1
CTT	5	6	1
TGT	6	2	1
TTC	7	1	1
CTC	8	1	1
CGC	1	23	0
TTT	2	26	0
CGT	3	14	0
TGC	4	5	0
CTT	5	5	0
TGT	6	3	0
TTC	7	2	0
CTC	8	1	0

Πίνακας 23: Δεδομένα της μελέτης "Common ABCB1 polymorphisms are not associated with multidrug resistance in epilepsy using a gene-wide tagging approach" (Leschziner, Andrew et al. 2007)

```
Logistic regression                                     Number of obs   =        149
                                                        LR chi2(7)      =      10.87
                                                        Prob > chi2     =     0.1442
Log likelihood = -97.569783                             Pseudo R2       =     0.0528

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   -.276753     .39915    -0.69   0.488    -1.059073    .5055667
_Ihaplotyp~3 |   -2.14262   .8066107    -2.66   0.008    -3.723548   -.5616925
_Ihaplotyp~4 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_Ihaplotyp~5 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_Ihaplotyp~6 |  -.6021754   .9552622    -0.63   0.528    -2.474455    1.270104
_Ihaplotyp~7 |  -.8898575   1.256659    -0.71   0.479    -3.352864    1.573149
_Ihaplotyp~8 |  -.1967103   1.441941    -0.14   0.891    -3.022862    2.629441
       _cons |   .1967103   .2814117     0.70   0.485    -.3548465    .7482671
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        149
                                                        LR chi2(7)      =      10.87
                                                        Prob > chi2     =     0.1442
Log likelihood = -97.569783                             Pseudo R2       =     0.0528

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   -.276753     .39915    -0.69   0.488    -1.059073    .5055667
_Ihaplotyp~3 |   -2.14262   .8066107    -2.66   0.008    -3.723548   -.5616925
_Ihaplotyp~4 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_Ihaplotyp~5 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_Ihaplotyp~6 |  -.6021754   .9552622    -0.63   0.528    -2.474455    1.270104
_Ihaplotyp~7 |  -.8898575   1.256659    -0.71   0.479    -3.352864    1.573149
_Ihaplotyp~8 |  -.1967103   1.441941    -0.14   0.891    -3.022862    2.629441
       _cons |   .1967103   .2814117     0.70   0.485    -.3548465    .7482671
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         16
                                                        LR chi2(15)     =     151.85
                                                        Prob > chi2     =     0.0000
Log likelihood = -27.939847                             Pseudo R2       =     0.7310

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .1967103   .2814117     0.70   0.485    -.3548465    .7482671
_Ihaplotyp~2 |   .1226023   .2862513     0.43   0.668    -.4384399    .6836445
_Ihaplotyp~3 |  -.4964369   .3389791    -1.46   0.143    -1.160824    .1679499
_Ihaplotyp~4 |  -1.526056   .4934352    -3.09   0.002    -2.493171   -.5589412
_Ihaplotyp~5 |  -1.526056   .4934352    -3.09   0.002    -2.493171   -.5589412
_Ihaplotyp~6 |  -2.036882   .6138498    -3.32   0.001    -3.240005   -.8337584
_Ihaplotyp~7 |  -2.442347   .7372098    -3.31   0.001    -3.887252   -.9974424
_Ihaplotyp~8 |  -3.135494   1.021508    -3.07   0.002    -5.137613   -1.133376
_IcasXhap_~2 |   -.276753     .39915    -0.69   0.488    -1.059073    .5055667
_IcasXhap_~3 |   -2.14262   .8066109    -2.66   0.008    -3.723549   -.5616922
_IcasXhap_~4 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_IcasXhap_~5 |  -.0143887   .6677269    -0.02   0.983    -1.323109    1.294332
_IcasXhap_~6 |  -.6021754   .9552622    -0.63   0.528    -2.474455    1.270104
_IcasXhap_~7 |  -.8898575   1.256659    -0.71   0.479    -3.352864    1.573149
_IcasXhap_~8 |  -.1967103   1.441941    -0.14   0.891    -3.022862    2.629441
       _cons |   3.135494   .2085144    15.04   0.000     2.726813    3.544175
```

Εικόνα 64: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Common ABCB1 polymorphisms are not associated with multidrug resistance in epilepsy using a gene-wide tagging approach" (Leschziner, Andrew et al. 2007)

haplotype	haplotype_nr	frequency	case
TTTCC	1	290	1
CCCAT	2	158	1
TTTCT	3	121	1
CTTAT	4	32	1
TTTCC	1	300	0
CCCAT	2	160	0
TTTCT	3	117	0
CTTAT	4	21	0

Πίνακας 24: Δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο ASPN

```
Logistic regression                                     Number of obs   =       1199
                                                        LR chi2(3)      =       2.54
                                                        Prob > chi2     =     0.4678
Log likelihood = -829.80897                             Pseudo R2       =     0.0015

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .0213228   .1391428     0.15   0.878    -.2513921    .2940377
_Ihaplotyp~3 |   .0675182   .1536004     0.44   0.660    -.2335331    .3685694
_Ihaplotyp~4 |    .455115   .2926612     1.56   0.120    -.1184904     1.02872
       _cons |  -.0339016   .0823505    -0.41   0.681    -.1953056    .1275025
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1199
                                                        LR chi2(3)      =       2.54
                                                        Prob > chi2     =     0.4678
Log likelihood = -829.80897                             Pseudo R2       =     0.0015

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .0213228   .1391428     0.15   0.878    -.2513921    .2940377
_Ihaplotyp~3 |   .0675182   .1536004     0.44   0.660    -.2335331    .3685694
_Ihaplotyp~4 |    .455115   .2926612     1.56   0.120    -.1184904     1.02872
       _cons |  -.0339016   .0823505    -0.41   0.681    -.1953056    .1275025
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =          8
                                                        LR chi2(7)      =     545.74
                                                        Prob > chi2     =     0.0000
Log likelihood = -26.150922                             Pseudo R2       =     0.9125

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.0339016   .0823505    -0.41   0.681    -.1953056    .1275025
_Ihaplotyp~2 |  -.6286087   .0978945    -6.42   0.000    -.8204784    -.436739
_Ihaplotyp~3 |  -.9416085    .108997    -8.64   0.000    -1.155239   -.7279784
_Ihaplotyp~4 |   -2.65926   .2257263   -11.78   0.000    -3.101676   -2.216845
_IcasXhap_~2 |   .0213228   .1391428     0.15   0.878    -.2513921    .2940377
_IcasXhap_~3 |   .0675182   .1536004     0.44   0.660    -.2335331    .3685694
_IcasXhap_~4 |    .455115   .2926613     1.56   0.120    -.1184907    1.028721
       _cons |   5.703782    .057735    98.79   0.000     5.590624    5.816941
------------------------------------------------------------------------------
```

Εικόνα 65: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο ASPN

haplotype	haplotype_nr	frequency	case
CC	1	408	1
TC	2	133	1
CT	3	60	1
TT	4	9	1
CC	1	390	0
TC	2	152	0
CT	3	57	0
TT	4	6	0

Πίνακας 25: Δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο COL2A1

```
Logistic regression                                     Number of obs   =       1215
                                                        LR chi2(3)      =       2.33
                                                        Prob > chi2     =     0.5060
Log likelihood =  -840.9965                             Pseudo R2       =     0.0014

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.1786518   .1382491    -1.29   0.196    -.4496152    .0923115
_Ihaplotyp~3 |   .0061729   .1980546     0.03   0.975    -.3820069    .3943527
_Ihaplotyp~4 |   .3603447   .5317826     0.68   0.498      -.68193    1.402619
       _cons |   .0451204   .0708173     0.64   0.524    -.0936788    .1839197
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1215
                                                        LR chi2(3)      =       2.33
                                                        Prob > chi2     =     0.5060
Log likelihood =  -840.9965                             Pseudo R2       =     0.0014

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.1786518   .1382491    -1.29   0.196    -.4496152    .0923115
_Ihaplotyp~3 |   .0061729   .1980546     0.03   0.975    -.3820069    .3943527
_Ihaplotyp~4 |   .3603447   .5317826     0.68   0.498      -.68193    1.402619
       _cons |   .0451204   .0708173     0.64   0.524    -.0936788    .1839197
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         8
                                                        LR chi2(7)      =   1194.13
                                                        Prob > chi2     =    0.0000
Log likelihood = -24.388095                             Pseudo R2       =    0.9608

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .0451204   .0708173     0.64   0.524    -.0936788    .1839197
_Ihaplotyp~2 |  -.9422662   .0956193    -9.85   0.000    -1.129677   -.7548558
_Ihaplotyp~3 |  -1.923095   .1418025   -13.56   0.000    -2.201023   -1.645168
_Ihaplotyp~4 |  -4.174387   .4113767   -10.15   0.000    -4.980671   -3.368104
_IcasXhap_~2 |  -.1786518   .1382491    -1.29   0.196    -.4496152    .0923115
_IcasXhap_~3 |   .0061729   .1980546     0.03   0.975    -.3820069    .3943527
_IcasXhap_~4 |   .3603447   .5317827     0.68   0.498    -.6819303     1.40262
       _cons |   5.966147    .050637   117.82   0.000       5.8669    6.065393
------------------------------------------------------------------------------
```

Εικόνα 66: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο COL2A1

haplotype	haplotype_nr	frequency	case
AC	1	492	1
AG	2	68	1
GC	3	44	1
AC	1	508	0
AG	2	59	0
GC	3	33	0

Πίνακας 26: Δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο COMP

```
Logistic regression                                     Number of obs   =       1204
                                                        LR chi2(2)      =       2.46
                                                        Prob > chi2     =     0.2926
Log likelihood = -833.31363                             Pseudo R2       =     0.0015

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |    .173973   .1888281     0.92   0.357    -.1961233    .5440693
_Ihaplotyp~3 |   .3196848   .2388123     1.34   0.181    -.1483787    .7877484
       _cons |  -.0320027   .0632537    -0.51   0.613    -.1559776    .0919721
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1204
                                                        LR chi2(2)      =       2.46
                                                        Prob > chi2     =     0.2926
Log likelihood = -833.31363                             Pseudo R2       =     0.0015

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |    .173973   .1888281     0.92   0.357    -.1961233    .5440693
_Ihaplotyp~3 |   .3196848   .2388123     1.34   0.181    -.1483787    .7877484
       _cons |  -.0320027   .0632537    -0.51   0.613    -.1559776    .0919721
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =          6
                                                        LR chi2(5)      =    1281.89
                                                        Prob > chi2     =     0.0000
Log likelihood = -19.524373                             Pseudo R2       =     0.9704

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Icase_1 |  -.0320027   .0632537    -0.51   0.613    -.1559776    .0919721
_Ihaplotyp~2 |  -2.152944   .1375415   -15.65   0.000     -2.42252   -1.883368
_Ihaplotyp~3 |  -2.733974   .1796428   -15.22   0.000    -3.086067    -2.38188
_IcasXhap_~2 |    .173973   .1888281     0.92   0.357    -.1961233    .5440693
_IcasXhap_~3 |   .3196848   .2388123     1.34   0.181    -.1483788    .7877484
       _cons |   6.230481   .0443678   140.43   0.000     6.143522    6.317441
------------------------------------------------------------------------------
```

Εικόνα 67: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο COMP

haplotype	haplotype_nr	frequency	case
CC	1	490	1
CG	2	38	1
TC	3	66	1
TG	4	30	1
CC	1	504	0
CG	2	34	0
TC	3	55	0
TG	4	24	0

Πίνακας 27: Δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο FRZB

```
Logistic regression                                     Number of obs   =       1241
                                                        LR chi2(3)      =       2.05
                                                        Prob > chi2     =     0.5622
Log likelihood = -859.15117                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .1393965   .2444432     0.57   0.569    -.3397034    .6184965
_Ihaplotyp~3 |   .2104924   .1932829     1.09   0.276    -.1683351    .5893199
_Ihaplotyp~4 |   .2513144   .2811138     0.89   0.371    -.2996584    .8022873
       _cons |  -.0281709   .0634424    -0.44   0.657    -.1525158     .096174
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1241
                                                        LR chi2(3)      =       2.05
                                                        Prob > chi2     =     0.5622
Log likelihood = -859.15117                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .1393965   .2444432     0.57   0.569    -.3397034    .6184965
_Ihaplotyp~3 |   .2104924   .1932829     1.09   0.276    -.1683351    .5893199
_Ihaplotyp~4 |   .2513144   .2811138     0.89   0.371    -.2996584    .8022873
       _cons |  -.0281709   .0634424    -0.44   0.657    -.1525158     .096174
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         8
                                                        LR chi2(7)      =   1689.80
                                                        Prob > chi2     =    0.0000
Log likelihood = -24.544099                             Pseudo R2       =    0.9718

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.0281709   .0634424    -0.44   0.657    -.1525158     .096174
_Ihaplotyp~2 |  -2.696216   .1771889   -15.22   0.000      -3.0435   -2.348932
_Ihaplotyp~3 |  -2.215243   .1420068   -15.60   0.000    -2.493571   -1.936915
_Ihaplotyp~4 |  -3.044522   .2089277   -14.57   0.000    -3.454013   -2.635032
_IcasXhap_~2 |   .1393965   .2444432     0.57   0.569    -.3397034    .6184965
_IcasXhap_~3 |   .2104924   .1932829     1.09   0.276    -.1683351    .5893199
_IcasXhap_~4 |   .2513144   .2811138     0.89   0.371    -.2996584    .8022873
       _cons |   6.222576   .0445435   139.70   0.000     6.135273     6.30988
------------------------------------------------------------------------------
```

Εικόνα 68: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Sex and Ethnic Differences in the Association of ASPN, CALM1, COL2A1, COMP, and FRZB With Genetic Susceptibility to Osteoarthritis of the Knee" (Valdes, Loughlin et al. 2007) – Για το γονίδιο FRZB

haplotype	haplotype_nr	frequency	case
ATAA	1	786	1
GGGG	2	646	1
GTGA	3	172	1
GGGA	4	171	1
GTAA	5	46	1
AGGG	6	36	1
ATGA	7	21	1
ATAA	1	756	0
GGGG	2	661	0
GTGA	3	170	0
GGGA	4	158	0
GTAA	5	48	0
AGGG	6	52	0
ATGA	7	31	0

Πίνακας 28: Δεδομένα της μελέτης "RGS4 is not a susceptibility gene for schizophrenia in Japanese:
Association study in a large case-control population" (Ishiguro, Horiuchi et al. 2007)

```
Logistic regression                                     Number of obs   =       3754
                                                        LR chi2(6)      =       6.18
                                                        Prob > chi2     =     0.4030
Log likelihood = -2598.9823                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.0618698   .0752055    -0.82   0.411    -.2092698    .0855303
_Ihaplotyp~3 |  -.0272194   .1195463    -0.23   0.820    -.2615259    .2070871
_Ihaplotyp~4 |   .0401531   .1215405     0.33   0.741    -.1980618     .278368
_Ihaplotyp~5 |   -.081475   .2125264    -0.38   0.701    -.4980192    .3350691
_Ihaplotyp~6 |  -.4066402   .2227185    -1.83   0.068    -.8431605    .0298801
_Ihaplotyp~7 |  -.4283802   .2871795    -1.49   0.136    -.9912418    .1344814
       _cons |   .0389154   .0509413     0.76   0.445    -.0609277    .1387585
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       3754
                                                        LR chi2(6)      =       6.18
                                                        Prob > chi2     =     0.4030
Log likelihood = -2598.9823                             Pseudo R2       =     0.0012

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.0618698   .0752055    -0.82   0.411    -.2092698    .0855303
_Ihaplotyp~3 |  -.0272194   .1195463    -0.23   0.820    -.2615259    .2070871
_Ihaplotyp~4 |   .0401531   .1215405     0.33   0.741    -.1980618     .278368
_Ihaplotyp~5 |   -.081475   .2125264    -0.38   0.701    -.4980192    .3350691
_Ihaplotyp~6 |  -.4066402   .2227185    -1.83   0.068    -.8431605    .0298801
_Ihaplotyp~7 |  -.4283802   .2871795    -1.49   0.136    -.9912418    .1344814
       _cons |   .0389154   .0509413     0.76   0.445    -.0609277    .1387585
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         14
                                                        LR chi2(13)     =    4074.67
                                                        Prob > chi2     =     0.0000
Log likelihood =  -47.11183                             Pseudo R2       =     0.9774

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .0389154   .0509413     0.76   0.445    -.0609277    .1387585
_Ihaplotyp~2 |  -.1342875   .0532505    -2.52   0.012    -.2386565   -.0299186
_Ihaplotyp~3 |  -1.492243   .0848829   -17.58   0.000     -1.65861   -1.325876
_Ihaplotyp~4 |  -1.565446   .0874749   -17.90   0.000    -1.736894   -1.393999
_Ihaplotyp~5 |   -2.75684   .1488492   -18.52   0.000    -3.048579   -2.465101
_Ihaplotyp~6 |  -2.676798    .143365   -18.67   0.000    -2.957788   -2.395807
_Ihaplotyp~7 |  -3.194054   .1832507   -17.43   0.000    -3.553219   -2.834889
_IcasXhap_~2 |  -.0618698   .0752055    -0.82   0.411    -.2092698    .0855303
_IcasXhap_~3 |  -.0272194   .1195463    -0.23   0.820    -.2615259    .2070871
_IcasXhap_~4 |   .0401531   .1215405     0.33   0.741    -.1980618     .278368
_IcasXhap_~5 |   -.081475   .2125264    -0.38   0.701    -.4980192    .3350691
_IcasXhap_~6 |  -.4066402   .2227186    -1.83   0.068    -.8431606    .0298802
_IcasXhap_~7 |  -.4283802   .2871796    -1.49   0.136    -.9912419    .1344815
       _cons |   6.628041   .0363696   182.24   0.000     6.556758    6.699325
------------------------------------------------------------------------------
```

Εικόνα 69: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "RGS4 is not a susceptibility gene for schizophrenia in Japanese: Association study in a large case-control population" (Ishiguro, Horiuchi et al. 2007)

haplotype	haplotype_nr	frequency	case
G T G T G C C G G A G C C C A T A	1	42	1
G C G T G C C G G A G C C G G C A	2	39	1
T T A C T T G G A A C C G C A T A	3	21	1
G T G T G C C G G A G C C C A C A	4	14	1
T T A C T T G G A A C C G C A T G	5	12	1
G C G T G C C G G A G T C G G C A	6	5	1
RARE	7	18	1
G T G T G C C G G A G C C C A T A	1	50	0
G C G T G C C G G A G C C G G C A	2	34	0
T T A C T T G G A A C C G C A T A	3	33	0
G T G T G C C G G A G C C C A C A	4	17	0
T T A C T T G G A A C C G C A T G	5	9	0
G C G T G C C G G A G T C G G C A	6	14	0
RARE	7	17	0

Πίνακας 29: Δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο MSH3

```
Logistic regression                                     Number of obs   =        325
                                                        LR chi2(6)      =       7.29
                                                        Prob > chi2     =     0.2951
Log likelihood = -220.81453                             Pseudo R2       =     0.0162

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .3115545   .3144238     0.99   0.322    -.3047048    .9278138
_Ihaplotyp~3 |  -.2776317   .3489006    -0.80   0.426    -.9614643    .4062008
_Ihaplotyp~4 |  -.0198026   .4172069    -0.05   0.962    -.8375132    .7979079
_Ihaplotyp~5 |   .4620355   .4881127     0.95   0.344    -.4946478    1.418719
_Ihaplotyp~6 |   -.855266   .5614606    -1.52   0.128    -1.955709    .2451766
_Ihaplotyp~7 |   .2315118   .3977293     0.58   0.561    -.5480233    1.011047
       _cons |  -.1743534   .2093072    -0.83   0.405    -.5845881    .2358813
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        325
                                                        LR chi2(6)      =       7.29
                                                        Prob > chi2     =     0.2951
Log likelihood = -220.81453                             Pseudo R2       =     0.0162

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .3115545   .3144238     0.99   0.322    -.3047048    .9278138
_Ihaplotyp~3 |  -.2776317   .3489006    -0.80   0.426    -.9614643    .4062008
_Ihaplotyp~4 |  -.0198026   .4172069    -0.05   0.962    -.8375132    .7979079
_Ihaplotyp~5 |   .4620355   .4881127     0.95   0.344    -.4946478    1.418719
_Ihaplotyp~6 |   -.855266   .5614606    -1.52   0.128    -1.955709    .2451766
_Ihaplotyp~7 |   .2315118   .3977293     0.58   0.561    -.5480233    1.011047
       _cons |  -.1743534   .2093072    -0.83   0.405    -.5845881    .2358813
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         14
                                                        LR chi2(13)     =     105.04
                                                        Prob > chi2     =     0.0000
Log likelihood = -33.692612                             Pseudo R2       =     0.6092

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.1743534   .2093072    -0.83   0.405    -.5845881    .2358813
_Ihaplotyp~2 |  -.3856625   .2222876    -1.73   0.083    -.8213381    .0500132
_Ihaplotyp~3 |  -.4155154   .2242834    -1.85   0.064    -.8551028    .0240719
_Ihaplotyp~4 |   -1.07881   .2807553    -3.84   0.000     -1.62908   -.5285394
_Ihaplotyp~5 |  -1.714798   .3620927    -4.74   0.000    -2.424487    -1.00511
_Ihaplotyp~6 |  -1.272966   .3023716    -4.21   0.000    -1.865603   -.6803283
_Ihaplotyp~7 |   -1.07881   .2807553    -3.84   0.000     -1.62908   -.5285394
_IcasXhap_~2 |   .3115545   .3144238     0.99   0.322    -.3047048    .9278138
_IcasXhap_~3 |  -.2776317   .3489006    -0.80   0.426    -.9614643    .4062008
_IcasXhap_~4 |  -.0198026   .4172069    -0.05   0.962    -.8375132    .7979079
_IcasXhap_~5 |   .4620355   .4881127     0.95   0.344    -.4946478    1.418719
_IcasXhap_~6 |   -.855266   .5614607    -1.52   0.128    -1.955709    .2451767
_IcasXhap_~7 |   .2315118   .3977293     0.58   0.561    -.5480233    1.011047
       _cons |   3.912023   .1414214    27.66   0.000     3.634842    4.189204
------------------------------------------------------------------------------
```

Εικόνα 70: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο MSH3

haplotype	haplotype_nr	frequency	case
A T C G A	1	82	1
A C T A C	2	27	1
A C C G A	3	32	1
C C C G A	4	11	1
A T C G A	1	89	0
A C T A C	2	43	0
A C C G A	3	33	0
C C C G A	4	5	0

Πίνακας 30: Δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο ERCC5

```
Logistic regression                                     Number of obs   =        322
                                                        LR chi2(3)      =       5.29
                                                        Prob > chi2     =     0.1517
Log likelihood = -220.04458                             Pseudo R2       =     0.0119

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.3834461   .2893509    -1.33   0.185    -.9505634    .1836712
_Ihaplotyp~3 |   .0511455   .2915203     0.18   0.861    -.5202239    .6225148
_Ihaplotyp~4 |   .8703745   .5606605     1.55   0.121    -.2284999    1.969249
       _cons |  -.0819171   .1530721    -0.54   0.593     -.381933    .2180987
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        322
                                                        LR chi2(3)      =       5.29
                                                        Prob > chi2     =     0.1517
Log likelihood = -220.04458                             Pseudo R2       =     0.0119

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.3834461   .2893509    -1.33   0.185    -.9505634    .1836712
_Ihaplotyp~3 |   .0511455   .2915203     0.18   0.861    -.5202239    .6225148
_Ihaplotyp~4 |   .8703745   .5606605     1.55   0.121    -.2284999    1.969249
       _cons |  -.0819171   .1530721    -0.54   0.593     -.381933    .2180987
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =          8
                                                        LR chi2(7)      =     164.89
                                                        Prob > chi2     =     0.0000
Log likelihood = -20.848817                             Pseudo R2       =     0.7982

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Icase_1 |  -.0819171   .1530721    -0.54   0.593     -.381933    .2180987
_Ihaplotyp~2 |  -.7274363   .1857196    -3.92   0.000     -1.09144   -.3634325
_Ihaplotyp~3 |  -.9921288   .2038112    -4.87   0.000    -1.391591   -.5926663
_Ihaplotyp~4 |  -2.879198   .4596041    -6.26   0.000    -3.780006   -1.978391
_IcasXhap_~2 |  -.3834461   .2893509    -1.33   0.185    -.9505634    .1836712
_IcasXhap_~3 |   .0511455   .2915203     0.18   0.861    -.5202239    .6225148
_IcasXhap_~4 |   .8703745   .5606605     1.55   0.121    -.2284999    1.969249
       _cons |   4.488636   .1059998    42.35   0.000     4.280881    4.696392
------------------------------------------------------------------------------
```

Εικόνα 71: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο ERCC5

haplotype	haplotype_nr	frequency	case
TT G C C A T C	1	104	1
D– G C C A T C	2	15	1
TT A C C A C A	3	8	1
TT G G A G T C	4	11	1
RARE	5	14	1
TT G C C A T C	1	120	0
D– G C C A T C	2	15	0
TT A C C A C A	3	17	0
TT G G A G T C	4	12	0
RARE	5	7	0

Πίνακας 31: Δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο RPA3

```
Logistic regression                                     Number of obs   =        323
                                                        LR chi2(4)      =       5.76
                                                        Prob > chi2     =     0.2177
Log likelihood = -220.44664                             Pseudo R2       =     0.0129

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .1431008   .3889499     0.37   0.713      -.619227    .9054287
_Ihaplotyp~3 |   -.610671   .4491907    -1.36   0.174    -1.491068    .2697266
_Ihaplotyp~4 |   .0560895   .4383961     0.13   0.898    -.8031511      .91533
_Ihaplotyp~5 |    .836248   .4819071     1.74   0.083    -.1082725    1.780769
       _cons |  -.1431008   .1339728    -1.07   0.285    -.4056828    .1194811
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        323
                                                        LR chi2(4)      =       5.76
                                                        Prob > chi2     =     0.2177
Log likelihood = -220.44664                             Pseudo R2       =     0.0129

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .1431008   .3889499     0.37   0.713      -.619227    .9054287
_Ihaplotyp~3 |   -.610671   .4491907    -1.36   0.174    -1.491068    .2697266
_Ihaplotyp~4 |   .0560895   .4383961     0.13   0.898    -.8031511      .91533
_Ihaplotyp~5 |    .836248   .4819071     1.74   0.083    -.1082725    1.780769
       _cons |  -.1431008   .1339728    -1.07   0.285    -.4056828    .1194811
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         10
                                                        LR chi2(9)      =     375.75
                                                        Prob > chi2     =     0.0000
Log likelihood = -23.863871                             Pseudo R2       =     0.8873

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.1431008   .1339728    -1.07   0.285    -.4056828    .1194811
_Ihaplotyp~2 |  -2.079442   .2738613    -7.59   0.000      -2.6162   -1.542683
_Ihaplotyp~3 |  -1.954278   .2591464    -7.54   0.000    -2.462196   -1.446361
_Ihaplotyp~4 |  -2.302585    .302765    -7.61   0.000    -2.895994   -1.709177
_Ihaplotyp~5 |  -2.841582   .3888322    -7.31   0.000    -3.603679   -2.079484
_IcasXhap_~2 |   .1431008   .3889499     0.37   0.713      -.619227    .9054287
_IcasXhap_~3 |   -.610671   .4491907    -1.36   0.174    -1.491068    .2697266
_IcasXhap_~4 |   .0560895   .4383961     0.13   0.898    -.8031511      .91533
_IcasXhap_~5 |    .836248   .4819071     1.74   0.083    -.1082725    1.780769
       _cons |   4.787492   .0912871    52.44   0.000     4.608572    4.966411
------------------------------------------------------------------------------
```

Εικόνα 72: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers" (Michiels, Danoy et al. 2007) – Για το γονίδιο RPA3

haplotype	haplotype_nr	frequency	case
TTGTCC	1	674	1
CGAATT	2	111	1
TTGACC	3	30	1
TGAATT	4	33	1
CGATTT	5	12	1
CGAATC	6	9	1
23 others	7	44	1
TTGTCC	1	681	0
CGAATT	2	115	0
TTGACC	3	29	0
TGAATT	4	26	0
CGATTT	5	12	0
CGAATC	6	6	0
23 others	7	43	0

Πίνακας 32: Δεδομένα της μελέτης "Haplotypic variation in MRE11, RAD50 and NBS1 and risk of non-Hodgkin's lymphoma" (Rollinson, Kesby et al. 2006)

```
Logistic regression                                     Number of obs   =       1825
                                                        LR chi2(6)      =       1.57
                                                        Prob > chi2     =     0.9546
Log likelihood = -1264.2076                             Pseudo R2       =     0.0006

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.0250697   .1437246    -0.17   0.862    -.3067648    .2566254
_Ihaplotyp~3 |   .0442337   .2660229     0.17   0.868    -.4771616    .5656291
_Ihaplotyp~4 |   .2487432   .2677997     0.93   0.353    -.2761345     .773621
_Ihaplotyp~5 |   .0103322    .411848     0.03   0.980    -.7968751    .8175394
_Ihaplotyp~6 |   .4157973   .5298393     0.78   0.433    -.6226687    1.454263
_Ihaplotyp~7 |   .0333217    .221213     0.15   0.880    -.4002478    .4668912
       _cons |  -.0103322   .0543333    -0.19   0.849    -.1168235    .0961591
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1825
                                                        LR chi2(6)      =       1.57
                                                        Prob > chi2     =     0.9546
Log likelihood = -1264.2076                             Pseudo R2       =     0.0006

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.0250697   .1437246    -0.17   0.862    -.3067648    .2566254
_Ihaplotyp~3 |   .0442337   .2660229     0.17   0.868    -.4771616    .5656291
_Ihaplotyp~4 |   .2487432   .2677997     0.93   0.353    -.2761345     .773621
_Ihaplotyp~5 |   .0103322    .411848     0.03   0.980    -.7968751    .8175394
_Ihaplotyp~6 |   .4157973   .5298393     0.78   0.433    -.6226687    1.454263
_Ihaplotyp~7 |   .0333217    .221213     0.15   0.880    -.4002478    .4668912
       _cons |  -.0103322   .0543333    -0.19   0.849    -.1168235    .0961591
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         14
                                                        LR chi2(13)     =    3661.63
                                                        Prob > chi2     =     0.0000
Log likelihood = -39.178361                             Pseudo R2       =     0.9790

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.0103322   .0543333    -0.19   0.849    -.1168235    .0961591
_Ihaplotyp~2 |   -1.77863   .1008171   -17.64   0.000    -1.976228   -1.581032
_Ihaplotyp~3 |  -3.156266    .189608   -16.65   0.000    -3.527891   -2.784642
_Ihaplotyp~4 |  -3.265466   .1998248   -16.34   0.000    -3.657115   -2.873816
_Ihaplotyp~5 |  -4.038656   .2912074   -13.87   0.000    -4.609412     -3.4679
_Ihaplotyp~6 |  -4.731803   .4100428   -11.54   0.000    -5.535472   -3.928134
_Ihaplotyp~7 |  -2.762362   .1572394   -17.57   0.000    -3.070546   -2.454179
_IcasXhap_~2 |  -.0250697   .1437246    -0.17   0.862    -.3067648    .2566254
_IcasXhap_~3 |   .0442337   .2660229     0.17   0.868    -.4771616    .5656291
_IcasXhap_~4 |   .2487432   .2677997     0.93   0.353    -.2761345     .773621
_IcasXhap_~5 |   .0103322    .411848     0.03   0.980    -.7968751    .8175394
_IcasXhap_~6 |   .4157973   .5298395     0.78   0.433     -.622669    1.454264
_IcasXhap_~7 |   .0333217    .221213     0.15   0.880    -.4002478    .4668912
       _cons |   6.523562   .0383201   170.24   0.000     6.448456    6.598668
------------------------------------------------------------------------------
```

Εικόνα 73: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Haplotypic variation in MRE11, RAD50 and NBS1 and risk of non-Hodgkin's lymphoma" (Rollinson, Kesby et al. 2006)

haplotype	haplotype_nr	frequency	case
AACGCCT	1	295	1
AACGCTC	2	161	1
AGCACCC	3	105	1
AACGCCC	4	35	1
TACGCCT	5	42	1
AACGGTC	6	35	1
Rare	7	21	1
AACGCCT	1	640	0
AACGCTC	2	359	0
AGCACCC	3	219	0
AACGCCC	4	109	0
TACGCCT	5	94	0
AACGGTC	6	62	0
Rare	7	62	0

Πίνακας 33: Δεδομένα της μελέτης "Effect of ATM, CHEK2 and ERBB2 TAGSNPs and haplotypes on endometrial cancer risk" (Einarsdottir, Humphreys et al. 2007) – Για το γονίδιο ATM

```
Logistic regression                                     Number of obs   =       2239
                                                        LR chi2(6)      =       6.04
                                                        Prob > chi2     =     0.4192
Log likelihood = -1383.0763                             Pseudo R2       =     0.0022

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.0274252    .118106    -0.23   0.816    -.2589087    .2040583
_Ihaplotyp~3 |   .0393814    .137994     0.29   0.775    -.2310819    .3098448
_Ihaplotyp~4 |   -.361507   .2066351    -1.75   0.080    -.7665044    .0434904
_Ihaplotyp~5 |  -.0311323   .1984947    -0.16   0.875    -.4201748    .3579102
_Ihaplotyp~6 |   .2027065   .2228291     0.91   0.363    -.2340304    .6394434
_Ihaplotyp~7 |  -.3081191   .2621076    -1.18   0.240    -.8218406    .2056024
       _cons |  -.7744928   .0703728   -11.01   0.000     -.912421   -.6365647
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       2239
                                                        LR chi2(6)      =       6.04
                                                        Prob > chi2     =     0.4192
Log likelihood = -1383.0763                             Pseudo R2       =     0.0022

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.0274252    .118106    -0.23   0.816    -.2589087    .2040583
_Ihaplotyp~3 |   .0393814    .137994     0.29   0.775    -.2310819    .3098448
_Ihaplotyp~4 |   -.361507   .2066351    -1.75   0.080    -.7665044    .0434904
_Ihaplotyp~5 |  -.0311323   .1984947    -0.16   0.875    -.4201748    .3579102
_Ihaplotyp~6 |   .2027065   .2228291     0.91   0.363    -.2340304    .6394434
_Ihaplotyp~7 |  -.3081191   .2621076    -1.18   0.240    -.8218406    .2056024
       _cons |  -.7744928   .0703728   -11.01   0.000     -.912421   -.6365647
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         14
                                                        LR chi2(13)     =    1939.51
                                                        Prob > chi2     =     0.0000
Log likelihood =  -45.15153                             Pseudo R2       =     0.9555

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.7744928   .0703728   -11.01   0.000     -.912421   -.6365647
_Ihaplotyp~2 |  -.5781458   .0659395    -8.77   0.000    -.7073848   -.4489068
_Ihaplotyp~3 |  -1.072396   .0782861   -13.70   0.000    -1.225834   -.9189586
_Ihaplotyp~4 |   -1.77012   .1036186   -17.08   0.000    -1.973209   -1.567032
_Ihaplotyp~5 |  -1.918173   .1104572   -17.37   0.000    -2.134666   -1.701681
_Ihaplotyp~6 |  -2.334334   .1330095   -17.55   0.000    -2.595028    -2.07364
_Ihaplotyp~7 |  -2.334334   .1330095   -17.55   0.000    -2.595028    -2.07364
_IcasXhap_~2 |  -.0274252    .118106    -0.23   0.816    -.2589087    .2040583
_IcasXhap_~3 |   .0393814    .137994     0.29   0.775    -.2310819    .3098448
_IcasXhap_~4 |   -.361507   .2066351    -1.75   0.080    -.7665044    .0434904
_IcasXhap_~5 |  -.0311323   .1984947    -0.16   0.875    -.4201748    .3579102
_IcasXhap_~6 |   .2027065   .2228291     0.91   0.363    -.2340304    .6394434
_IcasXhap_~7 |  -.3081191   .2621076    -1.18   0.240    -.8218406    .2056024
       _cons |   6.461468   .0395285   163.46   0.000     6.383994    6.538943
------------------------------------------------------------------------------
```

Εικόνα 74: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Effect of ATM, CHEK2 and ERBB2 TAGSNPs and haplotypes on endometrial cancer risk" (Einarsdottir, Humphreys et al. 2007) – Για το γονίδιο ATM

haplotype	haplotype_nr	frequency	case
GCCCCC	1	154	1
GGCTGC	2	161	1
GCCCCG	3	91	1
ACCCGC	4	91	1
GCTCGG	5	56	1
GGCCGC	6	35	1
Rare	7	105	1
GCCCCC	1	375	0
GGCTGC	2	359	0
GCCCCG	3	203	0
ACCCGC	4	156	0
GCTCGG	5	125	0
GGCCGC	6	94	0
Rare	7	234	0

Πίνακας 34: Δεδομένα της μελέτης "Effect of ATM, CHEK2 and ERBB2 TAGSNPs and haplotypes on endometrial cancer risk" (Einarsdottir, Humphreys et al. 2007) – Για το γονίδιο CHEK2

```
Logistic regression                                     Number of obs   =       2239
                                                        LR chi2(6)      =       5.64
                                                        Prob > chi2     =     0.4651
Log likelihood = -1382.4747                             Pseudo R2       =     0.0020

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .0880554   .1347474     0.65   0.513    -.1760447    .3521555
_Ihaplotyp~3 |    .087627   .1583518     0.55   0.580    -.2227369    .3979908
_Ihaplotyp~4 |   .3509769   .1629707     2.15   0.031     .0315603    .6703936
_Ihaplotyp~5 |   .0870114   .1871291     0.46   0.642     -.279755    .4537778
_Ihaplotyp~6 |  -.0979733   .2199316    -0.45   0.656    -.5290313    .3330847
_Ihaplotyp~7 |   .0886127   .1515173     0.58   0.559    -.2083558    .3855811
       _cons |  -.8899734   .0957088    -9.30   0.000    -1.077559   -.7023876
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       2239
                                                        LR chi2(6)      =       5.64
                                                        Prob > chi2     =     0.4651
Log likelihood = -1382.4747                             Pseudo R2       =     0.0020

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .0880554   .1347474     0.65   0.513    -.1760447    .3521555
_Ihaplotyp~3 |    .087627   .1583518     0.55   0.580    -.2227369    .3979908
_Ihaplotyp~4 |   .3509769   .1629707     2.15   0.031     .0315603    .6703936
_Ihaplotyp~5 |   .0870114   .1871291     0.46   0.642     -.279755    .4537778
_Ihaplotyp~6 |  -.0979733   .2199316    -0.45   0.656    -.5290313    .3330847
_Ihaplotyp~7 |   .0886127   .1515173     0.58   0.559    -.2083558    .3855811
       _cons |  -.8899734   .0957088    -9.30   0.000    -1.077559   -.7023876
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         14
                                                        LR chi2(13)     =     798.41
                                                        Prob > chi2     =     0.0000
Log likelihood = -47.063489                             Pseudo R2       =     0.8945

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -.8899734   .0957088    -9.30   0.000    -1.077559   -.7023876
_Ihaplotyp~2 |  -.0436036   .0738389    -0.59   0.555    -.1883252    .1011179
_Ihaplotyp~3 |   -.61372   .0871365    -7.04   0.000    -.7845045   -.4429356
_Ihaplotyp~4 |    -.87707   .0952729    -9.21   0.000    -1.063801   -.6903386
_Ihaplotyp~5 |  -1.098612   .1032796   -10.64   0.000    -1.301036   -.8961881
_Ihaplotyp~6 |  -1.383631   .1153471   -12.00   0.000    -1.609707   -1.157555
_Ihaplotyp~7 |  -.4716049   .0833077    -5.66   0.000     -.634885   -.3083248
_IcasXhap_~2 |   .0880554   .1347474     0.65   0.513    -.1760447    .3521555
_IcasXhap_~3 |    .087627   .1583518     0.55   0.580    -.2227369    .3979908
_IcasXhap_~4 |   .3509769   .1629707     2.15   0.031     .0315603    .6703936
_IcasXhap_~5 |   .0870114   .1871291     0.46   0.642     -.279755    .4537778
_IcasXhap_~6 |  -.0979733   .2199316    -0.45   0.656    -.5290313    .3330847
_IcasXhap_~7 |   .0886127   .1515173     0.58   0.559    -.2083558    .3855811
       _cons |   5.926926   .0516398   114.77   0.000     5.825714    6.028138
------------------------------------------------------------------------------
```

Εικόνα 75: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Effect of ATM, CHEK2 and ERBB2 TAGSNPs and haplotypes on endometrial cancer risk" (Einarsdottir, Humphreys et al. 2007) – Για το γονίδιο CHEK2

haplotype	haplotype_nr	frequency	case
CGTGGT	1	68	1
TGTAAC	2	45	1
TATAAC	3	25	1
TGGAGC	4	27	1
CGTGGT	1	68	0
TGTAAC	2	46	0
TATAAC	3	37	0
TGGAGC	4	21	0

Πίνακας 35: Δεδομένα της μελέτης "The Mannose-binding lectin (MBL2) haplotype and breast cancer: an association study in African American and Caucasian women" (Bernig, Boersma et al. 2007)

```
Logistic regression                                     Number of obs   =        337
                                                        LR chi2(3)      =       2.95
                                                        Prob > chi2     =     0.3986
Log likelihood = -232.04047                             Pseudo R2       =     0.0063

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.0219789   .2708747    -0.08   0.935    -.5528836    .5089258
_Ihaplotyp~3 |  -.3920421   .3105459    -1.26   0.207    -1.000701    .2166167
_Ihaplotyp~4 |   .2513144   .3377393     0.74   0.457    -.4106425    .9132713
       _cons |  -2.87e-16   .1714986    -0.00   1.000    -.3361311    .3361311
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =        337
                                                        LR chi2(3)      =       2.95
                                                        Prob > chi2     =     0.3986
Log likelihood = -232.04047                             Pseudo R2       =     0.0063

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.0219789   .2708747    -0.08   0.935    -.5528836    .5089258
_Ihaplotyp~3 |  -.3920421   .3105459    -1.26   0.207    -1.000701    .2166167
_Ihaplotyp~4 |   .2513144   .3377393     0.74   0.457    -.4106425    .9132713
       _cons |   1.78e-16   .1714986     0.00   1.000    -.3361311    .3361311
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =          8
                                                        LR chi2(7)      =      55.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -21.992498                             Pseudo R2       =     0.5572

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |  -1.60e-16   .1714986    -0.00   1.000    -.3361311    .3361311
_Ihaplotyp~2 |  -.3908663   .1909058    -2.05   0.041    -.7650347   -.0166979
_Ihaplotyp~3 |  -.6085898   .2042863    -2.98   0.003    -1.008984   -.2081959
_Ihaplotyp~4 |  -1.174985   .2496496    -4.71   0.000     -1.66429    -.685681
_IcasXhap_~2 |  -.0219789   .2708747    -0.08   0.935    -.5528836    .5089258
_IcasXhap_~3 |  -.3920421    .310546    -1.26   0.207    -1.000701    .2166168
_IcasXhap_~4 |   .2513144   .3377393     0.74   0.457    -.4106425    .9132713
       _cons |   4.219508   .1212678    34.79   0.000     3.981827    4.457188
------------------------------------------------------------------------------
```

Εικόνα 76: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "The Mannose-binding lectin (MBL2) haplotype and breast cancer: an association study in African American and Caucasian women" (Bernig, Boersma et al. 2007)

haplotype	haplotype_nr	frequency	case
G G G C A	1	1147	1
G G T T G	2	298	1
G G G C G	3	163	1
A G G C A	4	125	1
G A G C A	5	97	1
G G G C A	1	529	0
G G T T G	2	132	0
G G G C G	3	73	0
A G G C A	4	51	0
G A G C A	5	44	0

**Πίνακας 36: Δεδομένα της μελέτης "Progesterone receptor polymorphisms and risk of breast cancer:
results from two Australian breast cancer studies" (Johnatty, Spurdle et al. 2008)**

```
Logistic regression                               Number of obs   =       2659
                                                  LR chi2(4)      =       0.57
                                                  Prob > chi2     =     0.9662
Log likelihood = -1649.6521                       Pseudo R2       =     0.0002

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |   .0403749   .1170199     0.35   0.730    -.1889799    .2697297
_Ihaplotyp~3 |   .0293741    .150319     0.20   0.845    -.2652457    .3239939
_Ihaplotyp~4 |   .1225714     .17427     0.70   0.482    -.2189915    .4641343
_Ihaplotyp~5 |   .0166047   .1892056     0.09   0.930    -.3542315    .3874408
       _cons |   .7739167   .0525566    14.73   0.000     .6709076    .8769258
------------------------------------------------------------------------------

Multinomial logistic regression                   Number of obs   =       2659
                                                  LR chi2(4)      =       0.57
                                                  Prob > chi2     =     0.9662
Log likelihood = -1649.6521                       Pseudo R2       =     0.0002

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |   .0403749   .1170199     0.35   0.730    -.1889799    .2697297
_Ihaplotyp~3 |   .0293741    .150319     0.20   0.845    -.2652457    .3239939
_Ihaplotyp~4 |   .1225714     .17427     0.70   0.482    -.2189915    .4641343
_Ihaplotyp~5 |   .0166047   .1892056     0.09   0.930    -.3542315    .3874408
       _cons |   .7739167   .0525566    14.73   0.000     .6709076    .8769258
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                Number of obs   =         10
                                                  LR chi2(9)      =    2904.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -34.396752                       Pseudo R2       =     0.9769

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icase_1 |   .7739167   .0525566    14.73   0.000     .6709076    .8769258
_Ihaplotyp~2 |  -1.388187    .097294   -14.27   0.000    -1.578879   -1.197494
_Ihaplotyp~3 |  -1.980529   .1248559   -15.86   0.000    -2.225242   -1.735816
_Ihaplotyp~4 |  -2.339163   .1466227   -15.95   0.000    -2.626538   -2.051788
_Ihaplotyp~5 |  -2.486799   .1569001   -15.85   0.000    -2.794317    -2.17928
_IcasXhap_~2 |   .0403749   .1170199     0.35   0.730    -.1889799    .2697297
_IcasXhap_~3 |   .0293741    .150319     0.20   0.845    -.2652457    .3239939
_IcasXhap_~4 |   .1225714     .17427     0.70   0.482    -.2189916    .4641344
_IcasXhap_~5 |   .0166047   .1892056     0.09   0.930    -.3542315    .3874408
       _cons |   6.270988   .0434783   144.23   0.000     6.185773    6.356204
------------------------------------------------------------------------------
```

**Εικόνα 77: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Progesterone receptor polymorphisms and risk of breast cancer:
results from two Australian breast cancer studies" (Johnatty, Spurdle et al. 2008)**

haplotype	haplotype_nr	frequency	case
G G C G A	1	163	1
G G A A A	2	11	1
G G C G G	3	52	1
G G A A G	4	49	1
G A A A G	5	10	1
G A A A A	6	25	1
G A A G G	7	23	1
G A A G A	8	12	1
G A C G A	9	57	1
A A A A G	10	46	1
A A A A A	11	98	1
G G C G A	1	180	0
G G A A A	2	17	0
G G C G G	3	47	0
G G A A G	4	51	0
G A A A G	5	21	0
G A A A A	6	27	0
G A A G G	7	15	0
G A A G A	8	16	0
G A C G A	9	64	0
A A A A G	10	75	0
A A A A A	11	116	0

Πίνακας 37: Δεδομένα της μελέτης "Cholesteryl Ester Transfer Protein (CETP) Genetic Variation and Early Onset of Non-Fatal Myocardial Infarction" (Meiner, Friedlander et al. 2008)

```
Logistic regression                                     Number of obs   =       1175
                                                        LR chi2(10)     =      11.84
                                                        Prob > chi2     =     0.2959
Log likelihood = -805.59415                             Pseudo R2       =     0.0073

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihaplotyp~2 |  -.3361114    .401775    -0.84   0.403    -1.123576    .4513531
_Ihaplotyp~3 |   .2003028   .2284686     0.88   0.381    -.2474874    .6480929
_Ihaplotyp~4 |   .0592013   .2273907     0.26   0.795    -.3864763    .5048789
_Ihaplotyp~5 |  -.6427307    .399136    -1.61   0.107    -1.425023    .1395616
_Ihaplotyp~6 |   .0222456   .2978717     0.07   0.940    -.5615722    .6060635
_Ihaplotyp~7 |   .5266507   .3490494     1.51   0.131    -.1574735    1.210775
_Ihaplotyp~8 |  -.1884754   .3968928    -0.47   0.635    -.9663709    .5894201
_Ihaplotyp~9 |  -.0166252   .2118003    -0.08   0.937    -.4317462    .3984959
_Ihaploty~10 |  -.3896401   .2162475    -1.80   0.072    -.8134774    .0341973
_Ihaploty~11 |  -.0694161   .1746863    -0.40   0.691    -.4117949    .2729628
       _cons |  -.0992067   .1081227    -0.92   0.359    -.3111233     .11271
------------------------------------------------------------------------------

Multinomial logistic regression                         Number of obs   =       1175
                                                        LR chi2(10)     =      11.84
                                                        Prob > chi2     =     0.2959
Log likelihood = -805.59415                             Pseudo R2       =     0.0073

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
_Ihaplotyp~2 |  -.3361114    .401775    -0.84   0.403    -1.123576    .4513531
_Ihaplotyp~3 |   .2003028   .2284686     0.88   0.381    -.2474874    .6480929
_Ihaplotyp~4 |   .0592013   .2273907     0.26   0.795    -.3864763    .5048789
_Ihaplotyp~5 |  -.6427307    .399136    -1.61   0.107    -1.425023    .1395616
_Ihaplotyp~6 |   .0222456   .2978717     0.07   0.940    -.5615722    .6060635
_Ihaplotyp~7 |   .5266507   .3490494     1.51   0.131    -.1574735    1.210775
_Ihaplotyp~8 |  -.1884754   .3968928    -0.47   0.635    -.9663709    .5894201
_Ihaplotyp~9 |  -.0166252   .2118003    -0.08   0.937    -.4317462    .3984959
_Ihaploty~10 |  -.3896401   .2162475    -1.80   0.072    -.8134774    .0341973
_Ihaploty~11 |  -.0694161   .1746863    -0.40   0.691    -.4117949    .2729628
       _cons |  -.0992067   .1081227    -0.92   0.359    -.3111233     .11271
------------------------------------------------------------------------------
(case==0 is the base outcome)

Poisson regression                                      Number of obs   =         22
                                                        LR chi2(21)     =     767.41
                                                        Prob > chi2     =     0.0000
Log likelihood = -60.180276                             Pseudo R2       =     0.8644

------------------------------------------------------------------------------
   frequency |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Icase_1 |  -.0992067   .1081227    -0.92   0.359    -.3111233     .11271
_Ihaplotyp~2 |  -2.359744   .2537303    -9.30   0.000    -2.857046   -1.862441
_Ihaplotyp~3 |  -1.342809   .1638052    -8.20   0.000    -1.663862   -1.021757
_Ihaplotyp~4 |  -1.261131   .1586298    -7.95   0.000     -1.57204   -.9502226
_Ihaplotyp~5 |  -2.148434   .2305962    -9.32   0.000    -2.600395   -1.696474
_Ihaplotyp~6 |   -1.89712   .2063797    -9.19   0.000    -2.301617   -1.492623
_Ihaplotyp~7 |  -2.484907   .2687419    -9.25   0.000    -3.011631   -1.958182
_Ihaplotyp~8 |  -2.420368   .2608746    -9.28   0.000    -2.931673   -1.909063
_Ihaplotyp~9 |  -1.034074   .1455354    -7.11   0.000    -1.319318   -.7488296
_Ihaploty~10 |  -.8754687   .1374369    -6.37   0.000     -1.14484   -.6060975
_Ihaploty~11 |  -.4393667    .119064    -3.69   0.000    -.6727279   -.2060054
_IcasXhap_~2 |  -.3361114    .401775    -0.84   0.403    -1.123576    .4513531
_IcasXhap_~3 |   .2003028   .2284686     0.88   0.381    -.2474874    .6480929
_IcasXhap_~4 |   .0592013   .2273907     0.26   0.795    -.3864763    .5048789
_IcasXhap_~5 |  -.6427307    .399136    -1.61   0.107    -1.425023    .1395616
_IcasXhap_~6 |   .0222456   .2978717     0.07   0.940    -.5615722    .6060635
_IcasXhap_~7 |   .5266507   .3490494     1.51   0.131    -.1574735    1.210775
_IcasXhap_~8 |  -.1884754   .3968928    -0.47   0.635    -.9663709    .5894201
_IcasXhap_~9 |  -.0166252   .2118003    -0.08   0.937    -.4317462    .3984959
_IcasXhap~10 |  -.3896401   .2162475    -1.80   0.072    -.8134774    .0341973
_IcasXhap~11 |  -.0694161   .1746863    -0.40   0.691    -.4117949    .2729628
       _cons |   5.192957   .0745356    69.67   0.000      5.04687    5.339044
------------------------------------------------------------------------------
```

Εικόνα 78: Λογιστική, πολυωνιμική και Poisson παλινδρόμηση στα δεδομένα της μελέτης "Cholesteryl Ester Transfer Protein (CETP) Genetic Variation and Early Onset of Non-Fatal Myocardial Infarction" (Meiner, Friedlander et al. 2008)

ΒΙΒΛΙΟΓΡΑΦΙΑ

Agresti, A. (2002). "Categorical Data Analysis, 2nd edn.: John Wiley & Sons."

Baker, S. G. (2005). "A simple loglinear model for haplotype effects in a case-control study involving two unphased genotypes." Stat Appl Genet Mol Biol **4**: Article14.

Barrett, J. C. (2009). "Haploview: Visualization and analysis of SNP genotype data." CSH Protoc **2009**(10): pdb ip71.

Bauerfeind, A., H. Knoblauch, et al. (2002). "Single nucleotide polymorphism haplotypes in the cholesteryl-ester transfer protein (CETP) gene and lipid phenotypes." Hum Hered **54**(4): 166-73.

Berk, R. and J. MacDonald (2007). "Overdispersion and Poisson Regression." Journal of Quantitative Criminology **24**(3): 269-284.

Bernig, T., B. J. Boersma, et al. (2007). "The mannose-binding lectin (MBL2) haplotype and breast cancer: an association study in African-American and Caucasian women." Carcinogenesis **28**(4): 828-36.

Chen, H. Y. (2003). "A note on the prospective analysis of outcome-dependent samples." Journal of the Royal Statistical Society **65**(2): 575-584.

Chen, X. and Z. Li (2008). "Inference of haplotype effects in case-control studies using unphased genotype and environmental data." Biom J **50**(2): 270-82.

Chen, Y. H. and J. T. Kao (2006). "Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies." BMC Genet **7**: 43.

Clark, A. G. (1990). "Inference of haplotypes from PCR-amplified samples of diploid populations." Mol Biol Evol **7**(2): 111-22.

Clegg, M. T., J. F. Kidwell, et al. (1976). "Dynamics of correlated genetic systems. I. Selection in the region of the Glued locus of Drosophila melanogaster." Genetics **83**(4): 793-810.

Consonni, D., S. De Matteis, et al. (2010). "Lung cancer and occupation in a population-based case-control study." Am J Epidemiol **171**(3): 323-33.

Deloukas, P. and D. Bentley (2004). "The HapMap project and its application to genetic studies of drug response." Pharmacogenomics J **4**(2): 88-90.

Devlin, B. and N. Risch (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping." Genomics **29**(2): 311-22.

Eberly, L. E. (2007). "Correlation and simple linear regression." Methods Mol Biol **404**: 143-64.

Einarsdottir, K., K. Humphreys, et al. (2007). "Effect of ATM, CHEK2 and ERBB2 TAGSNPs and haplotypes on endometrial cancer risk." Hum Mol Genet **16**(2): 154-64.

Epstein, M. P. and G. A. Satten (2003). "Inference on haplotype effects in case-control studies using unphased genotype data." Am J Hum Genet **73**(6): 1316-29.

Fallin, D., A. Cohen, et al. (2001). "Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease." Genome Res **11**(1): 143-51.

French, B., T. Lumley, et al. (2006). "Simple estimates of haplotype relative risks in case-control data." Genet Epidemiol **30**(6): 485-94.

Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-9.

Gusfield, D. (2000). "A practical algorithm for optimal inference of haplotypes from diploid populations." Proc Int Conf Intell Syst Mol Biol **8**: 183-9.
Gusfield, D. (2001). "Inference of haplotypes from samples of diploid populations: complexity and algorithms." J Comput Biol **8**(3): 305-23.
HapMap (2003). "The International HapMap Project." Nature **426**(6968): 789-96.
HapMap (2004). "Integrating ethics and science in the International HapMap Project." Nat Rev Genet **5**(6): 467-75.
HapMap (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-320.
Hill, W. G. and B. S. Weir (1994). "Maximum-likelihood estimation of gene location by linkage disequilibrium." Am J Hum Genet **54**(4): 705-14.
Ishiguro, H., Y. Horiuchi, et al. (2007). "RGS4 is not a susceptibility gene for schizophrenia in Japanese: association study in a large case-control population." Schizophr Res **89**(1-3): 161-4.
Johnatty, S. E., A. B. Spurdle, et al. (2008). "Progesterone receptor polymorphisms and risk of breast cancer: results from two Australian breast cancer studies." Breast Cancer Res Treat **109**(1): 91-9.
Kankova, K., A. Stejskalova, et al. (2005). "Haplotype analysis of the RAGE gene: identification of a haplotype marker for diabetic nephropathy in type 2 diabetes mellitus." Nephrol Dial Transplant **20**(6): 1093-102.
Kankova, K., A. Stejskalova, et al. (2007). "Genetic risk factors for diabetic nephropathy on chromosomes 6p and 7q identified by the set-association approach." Diabetologia **50**(5): 990-9.
Kedda, M. A., D. L. Duffy, et al. (2006). "ADAM33 haplotypes are associated with asthma in a large Australian population." Eur J Hum Genet **14**(9): 1027-36.
Kehoe, P. G., H. Katzov, et al. (2003). "Haplotypes extending across ACE are associated with Alzheimer's disease." Hum Mol Genet **12**(8): 859-67.
Klerkx, A. H., M. W. Tanck, et al. (2003). "Haplotype analysis of the CETP gene: not TaqIB, but the closely linked -629C-->A polymorphism and a novel promoter variant are independently associated with CETP concentration." Hum Mol Genet **12**(2): 111-23.
Latiano, A., B. Mora, et al. (2007). "Analysis of candidate genes on chromosomes 5q and 19p in celiac disease." J Pediatr Gastroenterol Nutr **45**(2): 180-6.
Leschziner, G. D., T. Andrew, et al. (2007). "Common ABCB1 polymorphisms are not associated with multidrug resistance in epilepsy using a gene-wide tagging approach." Pharmacogenet Genomics **17**(3): 217-20.
Lin, C. P. and C. S. Fann (2009). "A novel tool for individual haplotype inference using mixed data." J Biomed Sci **16**: 52.
Lin, D. Y. and B. E. Huang (2007). "The Use of Inferred Haplotypes in Downstream Analyses." The American Journal of Human Genetics **80**(3): 577-579.
Lin, D. Y., D. Zeng, et al. (2005). "Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies." Genet Epidemiol **29**(4): 299-312.
Lin, S., D. J. Cutler, et al. (2002). "Haplotype inference in random population samples." Am J Hum Genet **71**(5): 1129-37.
Liu, Y., H. Zhang, et al. (2007). "Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma." Carcinogenesis **28**(9): 1906-13.
Lu, H., A. Inazu, et al. (2003). "Haplotype analyses of cholesteryl ester transfer protein gene promoter: a clue to an unsolved mystery of TaqIB polymorphism." J Mol Med **81**(4): 246-55.

Mander, A. P. (2001). "Haplotype analysis in population-based association studies." The Stata Journal **1**(1): 58–75.

Marchini, J., D. Cutler, et al. (2006). "A comparison of phasing algorithms for trios and unrelated individuals." Am J Hum Genet **78**(3): 437-50.

McCullagh, P. N., J. (1999). Generalized linear models.

Meiner, V., Y. Friedlander, et al. (2008). "Cholesteryl ester transfer protein (CETP) genetic variation and early onset of non-fatal myocardial infarction." Ann Hum Genet **72**(Pt 6): 732-41.

Michiels, S., P. Danoy, et al. (2007). "Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers." Carcinogenesis **28**(8): 1731-9.

Nei, M. and W. H. Li (1980). "Non-random association between electromorphs and inversion chromosomes in finite populations." Genet Res **35**(1): 65-83.

Niu, T. (2004). "Algorithms for inferring haplotypes." Genet Epidemiol **27**(4): 334-47.

Olson, J. M. and E. M. Wijsman (1994). "Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus." Am J Hum Genet **55**(3): 574-80.

Prentice, R. L. and R. Pyke (1979). "Logistic disease incidence models and case-control studies." Biometrika **66**(3): 403-411.

Qi, L., R. M. van Dam, et al. (2006). "Genetic variation in IL6 gene and type 2 diabetes: tagging-SNP haplotype analysis in large-scale case-control study and meta-analysis." Hum Mol Genet **15**(11): 1914-20.

Qian, L., J. Zhao, et al. (2007). "Brain-derived neurotrophic factor and risk of schizophrenia: an association study and meta-analysis." Biochem Biophys Res Commun **353**(3): 738-43.

Rollinson, S., H. Kesby, et al. (2006). "Haplotypic variation in MRE11, RAD50 and NBS1 and risk of non-Hodgkin's lymphoma." Leuk Lymphoma **47**(12): 2567-83.

Satten, G. A. and M. P. Epstein (2004). "Comparison of prospective and retrospective methods for haplotype inference in case-control studies." Genet Epidemiol **27**(3): 192-201.

Schaid, D. J., C. M. Rowland, et al. (2002). "Score tests for association between traits and haplotypes when linkage phase is ambiguous." Am J Hum Genet **70**(2): 425-34.

Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." Am J Hum Genet **73**(5): 1162-9.

Sun, S., C. M. Greenwood, et al. (2007). "Haplotype inference using a Bayesian Hidden Markov model." Genet Epidemiol **31**(8): 937-48.

Thomas, A. (2005). "Characterizing allelic associations from unphased diploid data by graphical modeling." Genet Epidemiol **29**(1): 23-35.

Tiret, L., P. Amouyel, et al. (1991). "Testing for association between disease and linked marker loci: a log-linear-model analysis." Am J Hum Genet **48**(5): 926-34.

Umbach, D. M. and C. R. Weinberg (1997). "Designing and analysing case-control studies to exploit independence of genotype and exposure." Stat Med **16**(15): 1731-43.

Valdes, A. M., J. Loughlin, et al. (2007). "Sex and ethnic differences in the association of ASPN, CALM1, COL2A1, COMP, and FRZB with genetic susceptibility to osteoarthritis of the knee." Arthritis Rheum **56**(1): 137-46.

Wallenstein, S., S. E. Hodge, et al. (1998). "Logistic regression model for analyzing extended haplotype data." Genet Epidemiol **15**(2): 173-81.

Wang, N., J. M. Akey, et al. (2002). "Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation." Am J Hum Genet **71**(5): 1227-34.

Weir, B. S. and S. R. Wilson (1986). "Log-linear models for linked loci." Biometrics **42**(3): 665-70.

Xing, E. P., M. I. Jordan, et al. (2007). "Bayesian haplotype inference via the Dirichlet process." J Comput Biol **14**(3): 267-84.

Xu, H., X. Wu, et al. (2004). "Comparison of haplotype inference methods using genotypic data from unrelated individuals." Hum Hered **58**(2): 63-8.

Zaykin, D. V., P. H. Westfall, et al. (2002). "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals." Hum Hered **53**(2): 79-91.

Zee, R. Y., N. R. Cook, et al. (2005). "Haplotype analysis of the beta2 adrenergic receptor gene and risk of myocardial infarction in humans." Genetics **169**(3): 1583-7.