ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ
Εργαστήριο Γενετικής Ποικιλότητας & Επιδημιολογίας

Εκπαιδευτικό Εργαστήριο Γενετικής Επιδημιολογίας

Στα πλαίσια της πραγματοποίησης συμποσίου για τη γενετική δομή των πληθυσμών και τις μελέτες συσχέτισης ολόκληρου του γονιδιώματος

Πρόγραμμα ΑΡΙΣΤΕΙΑ ΙΙ GENOMAP.GR

Γονιδιωματικός χάρτης αναφοράς της Ελλάδας. Μελέτη της δομής και ιστορίας των Ελληνικών υποπληθυσμών και της Ελληνικής διασποράς.

[Με συγχρηματοδότηση από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και εθνικούς πόρους στο πλαίσιο της πράξης ΑΡΙΣΤΕΙΑ ΙΙ (4386: GENOMAP.GR) του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Διά Βίου Μάθηση»]

Αλεξανδρούπολη, 27 Οκτωβρίου 2015

European Union
European Social Fund
OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY
Co-financed by Greece and the European Union
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

ΕΠΙΤΡΟΠΗ ΔΙΟΡΓΑΝΩΣΗΣ

Πάσχου Περιστέρα

Αναπληρώτρια Καθηγήτρια Γενετικής Πληθυσμών
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Μαντζάρης Δημήτριος

Μεταδιδακτορικός ερευνητής,
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Καραγιαννίδης Ιορδάνης

Υποψήφιος διδάκτορας
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Τσέτσος Φώτιος

Υποψήφιος διδάκτορας
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Καρατσώλη Μαρία

Προπτυχιακή φοιτήτρια,
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Τσιφιντάρης Μαργαρίτης

Προπτυχιακός φοιτητής
Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Μελέτες Συσχέτισης Ολόκληρου του Γονιδιώματος

Μια μελέτη συσχέτισης ολόκληρου του γονιδιώματος (**G**enome-**W**ide **A**ssociation **S**tudy – GWAS) είναι μια προσέγγιση που περιλαμβάνει σάρωση δεικτών από ολόκληρο το γονιδίωμα (για παράδεγμα ≈0,5 εκατομμύριο ή ένα εκατομμύριο) σε γονιδιωμάτα από πολλά άτομα (για παράδεγμα χιλιάδες ασθενείς και χιλιάδες άτομα ελέγχου) για την εύρεση γενετικών παραλλαγών (genetic variations) που σχετίζονται με μια ασθένεια.

Οι μελέτες συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Studies - GWAS) βασίζονται στην υπόθεση «κοινή παραλλαγή- κοινή ασθένεια» (common disease-common variant), δηλαδή οτι τα συχνά νοσήματα θα οφείλονται και σε παραλλαγές του γονιδιώματος που απαντώνται συχνά στον πληθυσμό. Επίσης, καθίστανται δυνατές χάρη στην δομή του γονιδιώματος και την ιδιότητα της ανισορροπίας σύνδεσης που παρατηρείται ανάμεσα στις παραλλαγές του DNA (μη τυχαία συσχέτιση αλληλομόρφων).

Με την ολοκλήρωση του **Human Genome Project** το **2003** και του **International HapMap Project** το **2005**, οι ερευνητές έχουν ένα σύνολο εργαλείων που καθιστούν δυνατή την διεξαγωγή μελετών GWAS. Τα εργαλεία περιλαμβάνουν ηλεκτρονικές βάσεις δεδομένων που περιέχουν την αλληλουχία του ανθρώπινου γονιδιώματος, χάρτες των ανθρώπινων γενετικών παραλλαγών και ένα σύνολο από νέες τεχνολογίες αλλά και αλγορίθμους που μπορούν με ταχύτητα και ακρίβεια να αναλύσουν δείγματα ολόκληρου του γονιδιώματος για γενετικές παραλλαγές που συντελούν στην εμφάνιση μιας νόσου. Αντικείμενο αυτού του εκπαιδευτικού εργαστηρίου αποτελεί η παρουσίαση εργαλείων που μπορούν να χρησιμοποιηθούν για την αναλυση δεδομένων μεγάλης κλίμακας στα πλαίσια GWAS με στόχο τον εντοπισμό γενετικών παραλλαγών που σχετίζονται με κοινά πολυπαραγονιτκά νοσήματα.

Μεθοδολογία των μελετών GWAS

Οι GWAS αφορούν μεγάλες μελέτες ασθενών και ατόμον ελέγχου (cases / controls). Απαιτούν μεγάλες ομάδες ατόμων, τα περισσότερα εκ των οποίων έχουν ένα συγκεκριμένο φαινότυπο μιας νόσου (cases), ενώ τα υπόλοιπα άτομα δεν παρουσιάζουν τη συγκεκριμένη νόσο (controls).

Γονοτύπηση κάθε ατόμου. Πρόκειται για την ανίχνευση ενός μεγάλου αριθμού SNPs που βρίσκονται σε ολόκληρο το γονιδίωμα. Ο αριθμός αυτός μπορεί να κυμαίνεται από 500.000 μέχρι και πάνω 1-2.000.000 δείκτες.

Ποιοτικός έλεγχος των δεδομένων που γονοτυπήθηκαν. Ενδεικτικά, η αφαίρεση ατόμων/SNPs με υψηλό ποσοστό ελλείψεων στα δεδομένα τους (missing data).

Αναζήτηση στατιστικών συσχετίσεων μεταξύ γονότυπων σε κάθε περιοχή και φαινοτυπικό τόπο για τον προσδιορισμό περιοχών που συνδέονται με την ευπάθεια στη νόσο. Μπορούν να χρησιμοποιηθούν SNP by SNP στατιστικοί έλεγχοι – x^2 ή παρόμοιοι.

Λεπτομερή χαρτογράφηση του σήματος συσχέτισης με άμεση γονοτύπηση επιπρόσθετων SNPs σε συναφείς περιοχές. Επίσης, λεπτομερή χαρτογράφηση της ανισορροπίας σύνδεσης (linkage disequilibrium) σε σχετικές περιοχές. Εμπειρική παραγωγή απλότυπων (χορδές του SNPs στο ίδιο χρωμόσωμα).

Αναπαραγωγή σε μια άλλη μεγάλη ανεξάρτητη ομάδα ασθενών (cases) και μαρτύρων (controls). Γονοτύπηση μόνον των SNPs που έχουν αποδειχθεί προηγουμένως, ότι συνδέονται με την ασθένεια που μελετάται. Αναπαραγωγή των αποτελεσμάτων χρησιμοποιώντας ελέγχους συσχέτισης.

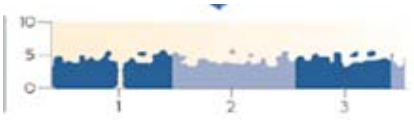
Βιολογική επικύρωση της συσχέτισης. Προσδιορισμός παραλλαγών που ενισχύουν τον κίνδυνο εκδήλωσης μιας ασθένειας, εξέταση λειτουργικών συνεπειών της παραλλαγής και καθορισμός μηχανισμού ενίσχυσης του κινδύνου.

Μεγάλος πληθυσμός με ασθενείς (cases) και μάρτυρες (Controls)
- Αποδοχή για παραμέτρους όπως η φυλή, η εθνικότητα και το φύλο
- Στρωματοποίηση ώστε να αυξηθούν τα σήματα

Γονοτύπηση SNPs Βασιζόμενη σε μικροσυστοιχίες
- ~1 εκατομμύριο τυχαίοι δείκτες SNPs
- ~25.000 SNPs που αυξάνουν τον κίνδυνο (π.χ. nsSNPs)

Παραγωγή απλοτύπων
- Βάση του International HapMap

Ανίχνευση σημάτων συσχέτισης
- Έλεγχος x^2 ή ανάλογη δοκιμασία
- Διόρθωση τιμών $P < 10^{-7}$



Χαρτογράφηση του σήματος συσχέτισης
- Κατευθυνόμενη γονοτύπηση πρόσθετων SNPs της μελετώμενης περιοχής
- Χαρτογράφηση της ανισορροπίας σύνδεσης στην περιοχή συσχέτισης
- Εμπειρική παραγωγή απλοτύπων
- Έλεγχος επίδρασης της πληθυσμιακής διαστρωμάτωσης, εάν είναι εφικτό

Επιβεβαίωση της συσχέτισης
- Μεγάλες ανεξάρτητες ομάδες Cases & Controls (n > 1000)
- Γονοτύπηση υποψήφιων SNPs (<20)
- Έλεγχος x^2 ή ανάλογη δοκιμασία για επιβεβαίωση του αρχικού σήματος

Γονότυποι	CC	AA	CA	Σύνολο
Αριθμός παρατηρούμενων ασθενών	59	27	98	184
Αριθμός παρατηρούμενων ατόμων ελέγχου	60	89	36	185
Σύνολο	119	116	134	369

Βιολογική Επαλήθευση της συσχέτισης
- Αναγνώριση παραλλαγής που αυξάνει τον κίνδυνο
- Έλεγχος του λειτουργικού αποτελέσματος της παραλλαγής
- Προσδιορισμός του μηχανισμού αύξησης του κινδύνου

(Kingsmore S, Lindquist I, Mudge J, Gessler D, Beavis W (2008), Genome-wide association studies: progress and potential for drug discovery and development, Nature Reviews Drug Discovery, 7, 221-230 | doi:10.1038/nrd2519)

Βιβλιογραφία

Paschou P, Drineas P, Yannaki E, Razou A, Kanaki K, Tsetsos F, Padmanabhuni S, Michalodimitrakis M, Renda M, Pavlovic R, Anagnostopoulos A, . Stamatoyannopoulos J, Kidd K, Stamatoyannopoulos G (2014), Maritime route of colonization of Europe, PNAS 111 (25), pp. 9211-9216.

Bush W, Moore J (2012), Genome-Wide Association Studies, PLOS Computational Biology, 90, pp. 7--24.

Stathias V, Sotiris G, Karagiannidis I, Bourikas G, Martinis G, Papazoglou D, Tavridou A, Papanas N, Maltezos E, Theodoridis M, Vargemezis V, Manolopoulos V, Speed W, Kidd J, Kidd K, Drineas P, Paschou P (2012), Exploring Genomic Structure Differences and imilarities between the Greek and European HapMap Populations: Implications for Association Studies, Annals of Human Genetics, 76, pp. 472-483

Visscher P, Brown A, McCarthy M, Yang J (2012), Five Years of GWAS Discovery, The American Journal of Human Genetics 90, pp. 7–24.

Paschou P, Lewis J, Javed A, Drineas P (2010), Ancestry informative markers for fine-scale individual assignment to worldwide populations, J Med Genet., 47, pp. 835-847

Kingsmore S, Lindquist I, Mudge J, Gessler D, Beavis W (2008), Genome-wide association studies: progress and potential for drug discovery and development, Nature Reviews Drug Discovery, 7, pp. 221-230.

McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, Ioannidis J, Hirschhorn J (2008), Genome-wide association studies for complex traits: consensus, uncertainty and challenges, Nature Publishing Group, 9, pp. 356-369

Pearson T, Manolio T (2008), How to Interpret a Genome-wide Association Study, The Journal of American Medical Association 299(11), pp. 1335-1344.

PLINK

```
@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|-----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2   |
|-----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:   |
|        http://pngu.mgh.harvard.edu/purcell/plink/         |
@----------------------------------------------------------@
```

1.ΕΙΣΑΓΩΓΗ

1.1 Τι είναι το PLINK

Το PLINK είναι ένα ελεύθερο λογισμικό και λογισμικό ανοιχτού κώδικα, για μελέτες ανάλυσης ολόκληρου του γονιδιώματος. Έχει σχεδιαστεί για να εκτελεί ένα εύρος βασικών, μεγάλης κλίμακας αναλύσεων με έναν υπολογιστικά αποδοτικό τρόπο. Το PLINK επικεντρώνεται στην ανάλυση δεδομένων γονοτύπου/φαινοτύπου. Πρόκειται για ένα πρόγραμμα γραμμής εντολών γραμμένο σε C/C++.

Το PLINK αναπτύχθηκε από τον Shaun Purcell στο Center for Human Genetic Research (CHGR), του Massachusetts General Hospital (MGH) σε συνεργασία με το Broad Institute of Harvard & MIT.

1.2 Χρήσεις του PLINK

Οι δυνατότητες του PLINK μπορούν να ταξινομηθούν σε οκτώ κατηγορίες, οι οποίες παρουσιάζονται στη συνέχεια.

1.2.1 Διαχείριση δεδομένων

- ✓ Ανάγνωση δεδομένων από ένα ευρύ φάσμα μορφών (formats).
- ✓ Αποκωδικοποίηση (recode) και επανατοποθέτηση (reorder) αρχείων.
- ✓ Ένωση (merge) δύο ή και περισσότερων αρχείων.
- ✓ Απομόνωση υποομάδων (SNPs ή ατόμων).
- ✓ Αντιστροφή της διεύθυνσης (flipping) κλώνων από SNPs.
- ✓ Συμπίεση δεδομένων σε αρχείο δυαδικής μορφής (binary files).

1.2.2 Περίληψη της στατιστικής (για ποιοτικό έλεγχο)

- ✓ Αλληλικές, γονοτυπικές συχνότητες, HWE έλεγχος.
- ✓ Ποσοστά απόντων γονοτύπων.
- ✓ Αιμομιξία, IBS και IBD στατιστική για άτομα μεμονωμένα και ζεύγη ατόμων.
- ✓ Μη-Μενδελική μεταβίβαση σε δεδομένα οικογενειών.
- ✓ Έλεγχος του φύλου βασισμένος σε SNPs του X χρωμοσώματος.

1.2.3 Ανίχνευση πληθυσμιακής διαστρωμάτωσης

- ✓ Ολοκληρωμένη σύνδεση ιεραρχικής στρωματοποίησης.
- ✓ Χειρισμός σχεδόν απεριόριστου αριθμού από SNPs.
- ✓ Πολυδιάστατη ανάλυση κλίμακας για οπτικοποίηση της υποδομής.
- ✓ Έλεγχος σημαντικότητας για το αν δύο άτομα ανήκουν ή όχι στον ίδιο πληθυσμό.

1.2.4 Βασικός έλεγχος συσχέτισης (association)

- ✓ Ασθενείς/Άτομα ελέγχου.
- ✓ Καθορισμένος αλληλικός έλεγχος.
- ✓ Επακριβής έλεγχος Fisher.
- ✓ Cochran-Armitage έλεγχος.
- ✓ Mantel-Haenszel και Breslow-Day έλεγχος για διαστρωματωμένα δείγματα.
- ✓ Επικρατή/Υπολειπόμενα και γενικά μοντέλα κληρονόμησης.

- ✓ Έλεγχοι σύγκρισης μοντέλων.
- ✓ Μελέτες βασισμένες στην οικογένεια.
- ✓ Ποσοτικά χαρακτηριστικά, συσχέτιση και αλληλεπίδραση.
- ✓ Συσχέτιση εξαρτώμενη από ένα ή περισσότερα SNPs.
- ✓ Ασυμπτωτικά και εμπειρικά p-values.
- ✓ Ευέλικτα προσαρμοσμένο σχήμα μεταλλαγών.

1.2.5. Μέσα πρόβλεψης πολλαπλών δεικτών, απλοτυπικά τεστ

- ✓ Ευέλικτα, υποθετικά απλοτυπικά τεστ.
- ✓ Συσχέτιση ασθενών/ατόμων ελέγχου και TDT στην πιθανολογική απλοτυπική φάση.
- ✓ Μία ομάδα από μεθόδους συσχέτισης για μελέτη μονού SNP σε συσχέτιση με το τοπικό απλοτυπικό πλαίσιο.
- ✓ Imputation, έλεγχος μη-τυποποιημένων SNPs δεδομένου ενός πάνελ αναφοράς.

1.2.6. Ανάλυση CNVs (Copy Number Variants)

- ✓ Σύνδεση SNPs και CNVs τεστ με στόχο την εύρεση CNVs.
- ✓ Φιλτράρισμα και περίληψη των διαδικασιών για σπάνια CNV δεδομένα.
- ✓ Τεστ σύγκρισης ασθενών/ατόμων ελέγχου για παγκόσμιες ιδιότητες CNV.
- ✓ Διαδικασία συσχέτισης βασισμένη σε μεταλλαγές για να ταυτοποιηθούν συγκεκριμένοι γενετικοί τόποι.

1.2.7. Επιπρόσθετα τεστ

- ✓ Τεστ συσχέτισης βασισμένα σε γονίδια.
- ✓ Έλεγχος για επίσταση.
- ✓ Αλληλεπίδραση γονιδίων-περιβάλλοντος σε ομοιόμορφα και διχοτομημένα περιβάλλοντα.

1.2.8. Μετά-ανάλυση

- ✓ Αυτόματα συνδυάζονται αρχεία γενετικά-μορφοποιημένα για εκατομμύρια SNPs.
- ✓ Ομαδοποίηση των αποτελεσμάτων στηριγμένη σε LD (Linkage Disequilibrium) και με βάση την περιοχή στηριγμένη σε πολλαπλές μελέτες.

2.ΤΥΠΟΙ ΑΡΧΕΙΩΝ ΤΟΥ PLINK

2.1 PED files

Τα PED αρχεία είναι white-space (space ή tab) οριοθετημένα αρχεία. Οι πρώτες έξι στήλες ενός PED αρχείου είναι υποχρεωτικές και παρουσιάζονται στη συνέχεια:

- ✓ Family ID
- ✓ Individual ID
- ✓ Paternal ID
- ✓ Maternal ID
- ✓ Sex (1=male, 2=female, other=unknown)
- ✓ Phenotype

Τα IDs είναι αλφαριθμητικά: ο συνδυασμός του ID της οικογένειας και του ατόμου θα πρέπει να ταυτοποιεί μοναδικά ένα άτομο. **Ένα PED αρχείο πρέπει να έχει ένα και μόνο ένα φαινότυπο στην έκτη στήλη.** Ο φαινότυπος μπορεί να είναι ένας ποσοτικός χαρακτήρας ή μία στήλη που θα δείχνει την κατάσταση επίδρασης. Η στήλη αυτή, εξ ορισμού, δηλώνεται ως εξής:

- ✓ -9 missing
- ✓ 0 missing
- ✓ 1 unaffected
- ✓ 2 affected

Σε ένα αρχείο PED μπορούν να προστεθούν σχόλια αφού προηγηθεί το σύμβολο της δίεσης (#).

2.2 MAP files

Εξ ορισμού κάθε γραμμή ενός MAP αρχείου περιγράφει ένα μόνο δείκτη και πρέπει να περιέχει ακριβώς τις ακόλουθες τέσσερεις στήλες:

- ✓ Chromosome (1-22, X, Y ή 0=unplaced)
- ✓ rs# ή snp identifier
- ✓ Genetic distance (Morgans)
- ✓ Base-pair position (bp units)

1 snp2 0 2
2 snp4 0 4
1 snp1 0 1
1 snp3 0 3
5 snp5 0 1

Το MAP αρχείο θα πρέπει να περιέχει τόσους δείκτες (markers) όσοι είναι και στο PED αρχείο.

2.3 Transposed filesets (Αντιμεταθετά σετ αρχείων)

Μία ακόμη μορφή αρχείων είναι αυτή των transposed (αντιμεταθετών) αρχείων, που περιλαμβάνουν δύο σετ αρχείων κειμένου: ένα TPED που περιλαμβάνει SNP και γενετική πληροφορία -όπου μία σειρά αντιστοιχεί σε ένα SNP- και ένα TFAM που περιλαμβάνει πληροφορίες για το άτομο και την οικογένεια - όπου μία σειρά αναφέρεται σε ένα άτομο.

Οι πρώτες 4 στήλες ενός TPED αρχείου είναι ίδιες με αυτές ενός MAP αρχείου με 4 στήλες. Έπειτα όλοι οι γονότυποι είναι σε λίστα για όλα τα άτομα και κάθε συγκεκριμένο SNP να αντιστοιχεί σε μία γραμμή. Το TFAM αρχείο περιλαμβάνει τις έξι πρώτες στήλες από ένα PED αρχείο. Καθένα από τα παρακάτω παραδείγματα PED/MAP αρχείων μπορούν να παρουσιαστούν σαν TPED/TFAM αρχεία.

<---- *normal.ped* ---->

```
1 1 0 0 1 1 A A G T
2 1 0 0 1 1 A C T G
3 1 0 0 1 1 C C G G
4 1 0 0 1 2 A C T T
5 1 0 0 1 2 C C G T
6 1 0 0 1 2 C C T T
```

<--- *normal.map* --->

```
1 snp1 0 5000650
1 snp2 0 5000830
```

<---------- *trans.tped* ---------->

```
1 snp1 0 5000650 A A A C C C A C C C C C
1 snp2 0 5000830 G T T G G G T T G T T T
```

<- *trans.tfam* ->

```
1 1 0 0 1 1
2 1 0 0 1 1
3 1 0 0 1 1
4 1 0 0 1 2
5 1 0 0 1 2
6 1 0 0 1 2
```

2.4 Δυαδικά PED αρχεία

Προκειμένου να εξοικονομηθεί χρόνος και χώρος, είναι δυνατόν να δημιουργηθεί ένα δυαδικό ped αρχείο (*.bed). Αυτό αποθηκεύει την πληροφορία για γενεαλογία/φαινότυπο σε ένα ξεχωριστό αρχείο και δημιουργεί ένα εκτεταμένο MAP αρχείο (*.bim), το οποίο περιέχει πληροφορία για τα ονόματα των αλληλίων. Για τη δημιουργία αυτών των αρχείων χρησιμοποιείται η εντολή:

plink –file mydata –make-bed

η οποία δημιουργεί (εξ ορισμού) τα τρία παρακάτω αρχεία
plink.bed (δυαδικό αρχείο, γονοτυπική πληροφορία)
plink.fam (οι πρώτες 6 στήλες από το mydata.ped)
plink.bim (εκτεταμένο MAP αρχείο: 2 επιπλέον στήλες= αλληλικά ονόματα)

Τα .fam και .bim αρχεία είναι απλά αρχεία κειμένου, γεγονός που σημαίνει ότι μπορούν να ανοίξουν με έναν απλό κειμενογράφο. Ωστόσο, η προβολή ενός .bed αρχείου δεν είναι εφικτή, διότι πρόκειται για ένα συμπιεσμένο, δυαδικό αρχείο.

Συνεπώς, για το χειρισμό ενός δυαδικού αρχείου, χρησιμοποιείται η επιλογή –bfile αντί της –file.

3.ΒΑΣΙΚΟΙ ΤΡΟΠΟΙ ΣΥΝΤΑΞΗΣ ΕΝΤΟΛΩΝ ΤΟΥ PLINK

Πληκτρολογώντας **plink** και προσδιορίζοντας ένα αρχείο χωρίς περαιτέρω επιλογές είναι δυνατόν να ελεγχθεί ένα αρχείο αναφορικά με την ακεραιότητά του. Επίσης, εμφανίζεται μια βασική περίληψη στατιστικών στοιχείων που αφορούν το αρχείο.

plink --file filename

Η επιλογή --file λαμβάνει μία μόνο παράμετρο, που είναι το όνομα του αρχείου εισόδου (input file) και ψάχνει για δύο αρχεία: ένα PED και ένα MAP με το ίδιο όνομα. Με άλλα λόγια, το --file filename δηλώνει τα filename.ped και filename.map αρχεία. Οι εντολές που εκτελούνται χρησιμοποιώντας το PLINK σώζονται σε ένα plink.log αρχείο.

3.1 Δυαδικό PED αρχείο

Το πρώτο βήμα είναι η δημιουργία ενός δυαδικού PED αρχείου. Αυτό είναι πιο συμπαγές στην παρουσίαση των δεδομένων, εξοικονομώντας χρόνο και επιταχύνοντας την επικείμενη ανάλυση. Για τη δημιουργία ενός δυαδικού PED αρχείου χρησιμοποιείται η ακόλουθη σύνταξη της εντολής plink:

plink --file filename --make-bed --out filename

Μετά την εκτέλεση της ανωτέρω ενέργειας, δημιουργούνται τα ακόλουθα τρία αρχεία: το δυαδικό αρχείο που περιέχει τα ακατέργαστα γονοτυπικά δεδομένα filename.bed καθώς επίσης και ένα αναθεωρημένο map αρχείο filename.bim το οποίο περιέχει δύο επιπλέον στήλες που δίνουν τα ονόματα των αλληλίων για κάθε SNP.Το τρίτο αρχείο που παράγεται είναι το filename.fam το οποίο περιλαμβάνει τις πρώτες έξι στήλες του filename.ped αρχείου. Είναι εφικτή η προβολή του περιεχομένου των .bim και .fam αρχείων όχι όμως του περιεχομένου του .bed.

3.2 Δουλεύοντας με δυαδικά αρχεία PED

Προκειμένου να προσδιοριστεί ότι τα προς επεξεργασία δεδομένα βρίσκονται με τη μορφή δυαδικού αρχείου, χρησιμοποιείται η επιλογή --bfile έναντι της --file.

plink --bfile filename

3.3 Στατιστική:Missing rates

Το PLINK δίνει τη δυνατότητα εξαγωγής απλών στατιστικών στοιχείων για τα ποσοστά εκλιπόντων δεδομένων στο αρχείο χρησιμοποιώντας την επιλογή --missing:

plink --bfile filename --missing --out miss_stat

Τα ανά άτομο και ανά SNP ποσοστά που προκύπτουν, εντοπίζονται στα αρχεία miss_stat.imiss και miss_stat.lmiss, αντίστοιχα. Αυτά τα αρχεία είναι απλά αρχεία κειμένου που μπορούν να προβληθούν με έναν απλό κειμενογράφο. Αν για παράδειγμα, χρειάζεται η προβολή του περιεχομένου του miss_stat.lmiss αρχείου πληκτρολογείται:

more miss_stat.lmiss

και προκύπτει:

CHR	SNP	N_MISS	F_MISS
1	rs6681049	0	0
1	rs4074137	0	0
1	rs7540009	0	0
1	rs1891905	0	0
1	rs9729550	0	0
1	rs3813196	0	0
1	rs6704013	2	0.0224719
1	rs307347	12	0.134831
1	rs9439440	2	0.0224719

Τα παραπάνω αποτελέσματα δείχνουν για κάθε SNP, τον αριθμό (N_MISS) και την αναλογία (F_MISS) των εκλιπόντων ατόμων.

Ομοίως, πληκτρολογώντας
more miss_stat.imiss

εμφανίζεται:

FID	IID	MISS_PHENO	N_MISS	F_MISS
HCB181	1	N	671	0.00803266
HCB182	1	N	1156	0.0138387
HCB183	1	N	498	0.00596164
HCB184	1	N	412	0.00493212
HCB185	1	N	329	0.00393852
HCB186	1	N	1233	0.0147605
HCB187	1	N	258	0.00308856

Η τελευταία στήλη είναι το πραγματικό γονοτυπικό ποσοστό για κάθε άτομο.

Επίσης, μπορεί να πραγματοποιηθεί ανάλυση των δεδομένων ανά χρωμόσωμα. Ενδεικτικά, στην περίπτωση ανάλυσης των δεδομένων αποκλειστικά για το χρωμόσωμα 1, τότε η εντολή που θα χρησιμοποιηθεί είναι:

plink --bfile filename --chr1 --out res1 --missing

3.4 Στατιστική: Αλληλικές συχνότητες

Η εντολή που ακολουθεί, παράγει ένα αρχείο που ονομάζεται freq_stat.frq το οποίο περιέχει τις αλληλικές συχνότητες και τους κωδικούς των αλληλίων για κάθε SNP.

plink --bfile filename --freq --out freq_stat

Στην περίπτωση που απαιτείται η γνώση της συχνότητας (Minor Allele Frequency – MAF) του σπανιότερου αλληλομόρφου ενός συγκεκριμένου SNP σε έναν πληθυσμό, μπορεί να χρησιμοποιηθεί η επιλογή --snp.

plink --bfile filename --snp snpname --freq --out freq_snpname

4.ΒΑΣΙΚΕΣ ΕΠΙΛΟΓΕΣ (OPTIONS) ΣΤΗΝ PLINK

Βασική είσοδος/έξοδος	Παράμετρος	Περιγραφή
--file	plink	Συγκεκριμενοποιεί .ped και .map αρχεία
--ped	plink.ped	Συγκεκριμενοποιεί .ped αρχεία
--map	plink.map	Συγκεκριμενοποιεί .map αρχεία
--no-sex		PED αρχείο που δεν περιέχει την 5η στήλη (sex)
--no-parents		PED αρχείο που δεν περιέχει τις στήλες 3,4 (parents)
--no-fid		PED αρχείο που δεν περιέχει την στήλη 1 (family ID)
--no-pheno		PED αρχείο που δεν περιέχει την στήλη 6 (phenotype)
--map3		Συγκεκριμενοποιεί την 3η στήλη του MAP αρχείου
--tfile	plink	Συγκεκριμενοποιεί .tped και .tfam αρχεία
--tped	plink.tped	Συγκεκριμενοποιεί .tped αρχεία
--tfam	plink.fam	Συγκεκριμενοποιεί .tfam αρχεία
--bfile	plink	Συγκεκριμενοποιεί .bed, .bed, και .fam
--bed	plink.bed	Συγκεκριμενοποιεί .bed αρχεία
--bim	plink.bim	Συγκεκριμενοποιεί .bim αρχεία
--fam	plink.fam	Συγκεκριμενοποιεί .fam αρχεία
--out	plink	Συγκεκριμενοποιεί output root filename.

Επιλογή SNPs και ατόμων	Παράμετρος	Περιγραφή
--chr	N	Επιλέγει ένα συγκεκριμένο N χρωμόσωμα
--gene	Name	Επιλέγει ένα συγκεκριμένο γονίδιο.
--from	SNP	Επιλέγει από ένα συγκεκριμένο SNP
--to	SNP	Έως ένα συγκεκριμένο SNP
--snps	SNP list	Επιλέγει SNPs που χωρίζονται με κόμμα το ένα από το άλλο
--snp	SNP	Επιλέγει ένα συγκεκριμένο SNP
--from-bp	bp	Επιλογή SNP
--to-bp	bp	Επιλογή SNP
--from-kb	kb	Επιλογή SNP
--to-kb	kb	Επιλογή SNP
--from-mb	mb	Επιλογή SNP
--to-mb	mb	Επιλογή SNP
--extract	snplist	Εξαγωγή λίστας SNPs.Δημιουργία αρχείου με τα συγκεκριμένα SNPs.
--exclude	snplist	Αποκλεισμός λίστας SNPs. Στο τελικό αρχείο δεν περιλαμβάνονται τα συγκεκριμένα SNPs που ζητήσαμε να εξαιρεθούν.
--keep	indlist	Κράτα μόνο συγκεκριμένα άτομα
--remove	indlist	Αφαίρεσε μόνο συγκεκριμένα άτομα
--keep-before-remove		Εκτέλεσε το keep πριν το remove
--exclude-before-extract		Εκτέλεσε το exclude πριν το extract
--filter	filename value	Φιλτράρισμα ατόμων
--mfilter	var #	Συγκεκριμενοποίηση του φίλτρου

--filter-cases		Φιλτράρισμα μόνο των ασθενών
--filter-controls		Φιλτράρισμα μόνο των ατόμων ελέγχου
--filter-males		Φιλτράρισμα αρσενικών ατόμων
--filter-females		Φιλτράρισμα θηλυκών ατόμων
--prune		Αφαίρεση ατόμων με εκλιπόντες φαινότυπους

Άλλες επιλογές για διαχείριση δεδομένων	**Περιγραφή**
--make-bed	Δημιουργία .bed, .fam .bim αρχείων
--recode	Παραγωγή νέων .ped και .map αρχείων
--merge	Ένωση σε ένα PED/MAP fileset
--bmerge	Ένωση σε ένα δυαδικό fileset

5.ΒΙΒΛΙΟΓΡΑΦΙΑ

Purchell S (2010), PLINK (1.07) Documentation.

EIGENSOFT SOFTWARE

1. ΕΙΣΑΓΩΓΗ

Το EIGENSTRAT ανιχνεύει και διορθώνει την πληθυσμιακή διαστρωμάτωση σε μελέτες συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Studies – GWAS). Η μέθοδος, που βασίζεται στην ανάλυση κύριων συνιστωσών (Principal Component Analysis – PCA), μοντελοποιεί διαφορές στην καταγωγή μεταξύ ασθενών (cases) και μαρτύρων (controls), κατά μήκος συνεχόμενων αξόνων όπου παρατηρούνται παραλλαγές. Η διόρθωση που προκύπτει είναι συγκεκριμένη για έναν υποψήφιο δείκτη παραλλαγής, βάση της συχνότητας των προγονικών πληθυσμών, ελαχιστοποιώντας εικονικές συσχετίσεις και μεγιστοποιώντας παράλληλα τη δύναμη ανίχνευσης πραγματικών συσχετίσεων. Η προσέγγιση είναι ισχυρή, καθώς και γρήγορη, με αποτέλεσμα να μπορεί να εφαρμοστεί σε μελέτες ασθενειών με εκατοντάδες χιλιάδες δεικτών.

Από το Δεκέμβριο του 2006, το EIGENSTRAT υλοποιείται σαν τμήμα του ολοκληρωμένου πακέτου EIGENSOFT. Ο πηγαίος κώδικας, τα εγχειρίδια και τα εκτελέσιμα αρχεία για το πακέτο EIGENSOFT είναι διαθέσιμα στην ιστοσελίδα του Reich.

2.ΤΥΠΟΙ ΑΡΧΕΙΩΝ

Η εφαρμογή υποστηρίζει πέντε διαφορετικούς τύπους αρχείων (file format). Ο όρος τύπος αρχείου αναφέρεται ταυτόχρονα σε τρία διακριτά αρχεία:

genotype file Περιλαμβάνει τα γονοτυπημένα δεδομένα για κάθε άτομο σε κάθε SNP

snp file Περιλαμβάνει πληροφορίες για κάθε SNP

indiv file Περιλαμβάνει πληροφορίες για κάθε άτομο

Το πακέτο EIGENSOFT δεν επιτρέπει περισσότερους από οκτώ δισεκατομμύρια γονότυπους, και θα αναφέρει ένα μήνυα σφάλματος εάν προσπαθήσει να δημιουργήσει ένα αρχείο εξόδου μεγαλύτερο από 2 GigaBytes.

2.1 ANCESTRYMAP format

Το αρχείο **genotype** περιλαμβάνει μια γραμμή για κάθε έγκυρο γονότυπο. Υπάρχουν οι εξής τρεις στήλες:

- ✓ 1η στήλη: Το όνομα του SNP.
- ✓ 2η στήλη: Το ID του δείγματος.
- ✓ 3η στήλη: Ο αριθμός του αλλήλιου αναφοράς (0 ή 1 ή 2).

Οι γονότυποι που έχουν χαθεί κωδικοποιούνται με την απουσία εισόδου στο αρχείο genotype.

Το αρχείο **snp** περιλαμβάνει μια γραμμή για κάθε SNP. Υπάρχουν οι εξής έξι στήλες:

- ✓ 1η στήλη: Το όνομα του SNP.
- ✓ 2η στήλη: Το χρωμόσωμα. Το X χρωμόσωμα κωδικοποιείται σαν 23, το Y χρωμόσωμα σαν 24, mtDNA σαν 90 και XY σαν 91. Το SNP με μη έγκυρες τιμές χρωμοσώματος, όπως μηδέν, θα αφαιρούνται.

- ✓ 3η στήλη: Γενετική θέση (σε Morgans). Εάν είναι άγνωστη, τότε τίθεται 0.0.
- ✓ 4η στήλη: Φυσική θέση (σε bases).
- ✓ 5η στήλη: Είναι το αλλήλιο αναφοράς.
- ✓ 6η στήλη: είναι η παραλλαγή του αλληλίου.

Το αρχείο indiv περιλαμβάνει μια γραμμή για κάθε άτομο. Υπάρχουν οι εξής τρεις στήλες:

- ✓ 1η στήλη: Το ID του δείγματος. Το μήκος περιορίζεται στους 39 χαρακτήρες, συμπεριλαμβανομένου του ονόματος της οικογένειας.
- ✓ 2η στήλη: Είναι το φύλο (M ή F). Εάν δεν είναι γνωστό, τότε τίθεται σε U (Unknown).
- ✓ 3η στήλη: Περιγράφει αν είναι ασθενής (case) ή μάρτυρας (control). Αν υπάρχει ο χαρακτηρισμός "Ignore", τότε το άτομο και τα γονοτυπικά δεδομένα του αφαιρούνται από το dataset σε όλα τα αρχεία που προκύπτουν.

Ο τύπος αρχείου ANCESTRYMAP χρησιμοποιείται για ιστορικούς λόγους. Το λογισμικό αυτό είναι πλήρως ανεξάρτητο από το λογισμικό 2004 ANCESTRYMAP.

2.2 EIGENSTRAT format

Το EIGENSTRAT format χρησιμοποιείται από το πρόγραμμα eigenstrat

Το αρχείο **genotype** περιλαμβάνει μια γραμμή για κάθε SNP. Κάθε γραμμή περιλαμβάνει ένα χαρακτήρα για κάθε άτομο, οποίος είναι:

- ✓ 0: Μηδενικά αντίγραφα του αλληλόμορφου αναφοράς.
- ✓ 1: Ένα αντίγραφο του αλληλόμορφου αναφοράς.
- ✓ 2: Δυο αντίγραφα του αλληλόμορφου αναφοράς.
- ✓ 9: Απώλεια δεδομένων.

2.3 PED format

Το αρχείο **genotype** έχει κατάληξη .ped και περιλαμβάνει μια γραμμή για κάθε άτομο. Κάθε γραμμή έχει έξι ή επτά στήλες με πληροφορίες σχετικά με το άτομο και δυο στήλες με τους γονότυπους για κάθε SNP, σύμφωνα με τη σειρά που προσδιορίζονται τα SNPs στο αρχείο snp. Η μορφή του γονότυπου **πρέπει** να είναι είτε **0ACGT** είτε **01234**, όπου μηδέν σημαίνει απολεσθέντα δεδομένα.

Οι πρώτες έξι ή επτά στήλες του αρχείου **genotype** είναι:

- ✓ 1η στήλη: Προσδιοριστικό της οικογένειας (family ID).
- ✓ 2η στήλη: Προσδιοριστικό του ατόμου (sample ID).
- ✓ 3η στήλη & 4η στήλη: Προσδιοριστικά των γονέων του ατόμου.
- ✓ 5η στήλη: Φύλο (1: άνδρας, 2: γυναίκα).
- ✓ 6η στήλη: case / control status (1: control, 2: case).

- ✓ 7η στήλη: Η στήλη είναι προαιρετική και έχει τιμή 1.

Το αρχείο **snp** περιλαμβάνει μια γραμμή για κάθε SNP. Υπάρχουν έξι στήλες (οι τελευταίες δυο προαιρετικές):

- ✓ 1η στήλη: Το χρωμόσωμα. Χρήση του X για το X χρωμόσωμα. Τα SNPs με μη έγκυρες τιμές χρωμοσώματος, όπως μηδέν, θα αφαιρούνται.
- ✓ 2η στήλη: Το όνομα του SNP.
- ✓ 3η στήλη: Γενετική θέση (σε Morgans).
- ✓ 4η στήλη: Φυσική θέση (σε bases).
- ✓ 5η & 6η στήλη: Είναι προαιρετικές και περιλαμβάνουν το αλληλόμορφο αναφοράς και την παραλλαγή του.

Το αρχείο **ndiv** περιλαμβάνει τις πρώτες έξι η επτά στήλες του αρχείου **genotype**.

2.4 PACKEDPED format

Το αρχείο **genotype** έχει κατάληξη .bed και να είναι σε SNP-major ταξινόμηση.

Το αρχείο **snp** έχει κατάληξη .pedsnp και πρέπει να είναι σε ταξινόμηση genomewide.

Το αρχείο **indiv** έχει κατάληξη .pedind.

2.5 PACKEDANCESTRYMAP format

Το αρχείο **genotype** έχει κατάληξη . packedancestrymapgeno.

Το αρχείο **snp** έχει κατάληξη . snp και πρέπει να είναι σε ταξινόμηση genomewide.

Το αρχείο **indiv** έχει κατάληξη .ind.

3 Smartpca

Το πρόγραμμα smartpca επιτρέπει την ανάλυση και τον υπολογισμό των κύριων συνιστωσών (Principal Component Analysis – PCA). Οι παράμετροι που μπορεί να δεχθεί το πρόγραμμα είναι οι ακόλουθες:

-i example.geno Το αρχείο genotype σε οποιοδήποτε format

-a example.snp Το αρχείο snp σε οποιοδήποτε format

-b example.ind Το αρχείο indiv σε οποιοδήποτε format

-k k Ο αριθμός των κύριων συνιστωσών που θα υπολογιστούν από το πρόγραμμα. Η προεπιλεγμένη τιμή είναι ίση με 10.

-o example.pca Το αρχείο εξόδου των κύριων συνιστωσών. Το αρχείο έχει κατάληξη .pca. Τα άτομα που έχουν αφαιρεθεί θεωρούμενα ως αποκλίνουσες τιμές (outliers) θα έχουν όλες τις τιμές τους ίσες με 0.0.

-p example.plot Πρόθεμα των αρχείων γραφικών παραστάσεων για τις δυο κορυφαίες κύριες συνιστώσες. Τα άτομα επισημαίνονται σύμφωνα με τις ετικέτες στο αρχείο indiv.

-e example.eval Το αρχείο εξόδου με όλες τις ιδιοτιμές. Το αρχείο έχει κατάληξη .eval.

-l example.log Αρχείο καταγραφής. Το αρχείο έχει κατάληξη .log.

-m maxiter Μέγιστος αριθμός επαναλήψεων για την αφαίρεση των αποκλινουσών τιμών (outliers). Η προεπιλεγμένη τιμή είναι 5, ενώ η απενεργοποίηση της παραμέτρου γίνεται με –m 0.

-t topk Ο αριθμός των κύριων συνιστωσών κατά τις οποίες αφαιρούνται οι αποκλίνουσες τιμές κατά τη διάρκεια κάθε επανάληψης αφαίρεσης των outliers. Η προεπιλεγμένη τιμή είναι 10.

-s sigma Ο αριθμός των τυπικών αποκλίσεων που πρέπει να υπερβεί ένα άτομο, από το σύνολο των κύριων συνιστωσών, προκειμένου να αφαιρεθεί σαν αποκλίνουσα τιμή (outlier).

Ο εκτιμώμενος χρόνος εκτέλεσης του προγράμματος smartpca είναι:

$2.5*10^{-12}$ *nSNP * $NSAMPLES^2$ (σε ώρες), εάν δεν αφαιρεθούν οι αποκλίνουσες τιμές (outliers)

$2.5*10^{-12}$ *nSNP * $NSAMPLES^2$ * (1-m) (σε ώρες), εάν γίνουν m επαναλήψεις αφαίρεσης των αποκλινουσών τιμών (outliers).

Στην περίπτωση που η μεταβλητή m λαμβάνει τιμές μικρότερες ή ίσες του 5, τότε ο εκτιμώμενος χρόνος εκτέλεσης του προγράμματος είναι μέχρι $1.5*10^{-11}$ * nSNP * $NSAMPLES^2$ σε ώρες.

Όπου nSNP είναι ο αριθμός των SNPs και NSAMPLES είναι ο αριθμός των ατόμων του συνόλου δεδομένων.

4. Βιβλιογραφία

Patterson N, Price A, Reich D (2006), Population structure and eigenanalysis. PLoS Genet 2(12): e190, pp. 2074-2093.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick M, Reich D (2006), Principal components analysis corrects for stratification in genome-wide association studies, Nature Genetics 38, pp. 904 - 909

HAPLOVIEW

1. Εισαγωγή

Το Haploview έχει σχεδιαστεί με στόχο την απλούστευση και την επιτάχυνση της διαδικασίας ανάλυσης απλότυπων παρέχοντας μια κοινή διεπαφή για διάφορες εργασίες που σχετίζονται με τέτοιου είδους αναλύσεις. Το Haploview είναι ένα ελεύθερο λογισμικό και λογισμικό ανοιχτού κώδικα.

2 Χρήσεις του Haploview

Το Haploview υποστηρίζει τις ακόλουθες λειτουργίες:

- ✓ Εκτίμηση της συχνότητας ενός απλότυπου για έναν πληθυσμό.
- ✓ Έλεγχοι συσχέτισης μεταξύ ενός SNPκαι ενός απλότυπου.
- ✓ Έλεγχος αντιμετάθεσης για τη σημαντικότητα της συσχέτισης.
- ✓ Εκτέλεση της εφαρμογής Tagger που δημιουργήθηκε από τον Paul de Bakker για την επιλογή και αξιολόγηση των tag SNPs.
- ✓ Αυτόματη λήψη των δεδομένων γονοτύπησης από το HapMap.
- ✓ Οπτικοποίηση και γραφική αναπαράσταση των αποτελεσμάτων του PLINK αναφορικά με μελέτες συσχέτισης ολόκληρου του γονιδιώματος συμπεριλαμβάνοντας προηγμένες επιλογές φιλτραρίσματος.

3. Βιβλιογραφία

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005 Jan 15 [PubMed ID: 15297300]

Information about the exact test for HW can be found in the following paper: Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005 May;76(5):887-93.

Information about parenTDT can be found in the following paper: Purcell S, Sham P, Daly MJ. Parental phenotypes in family-based association analysis. Am J Hum Genet. 2005 Feb;76(2):249-59.

1. Login

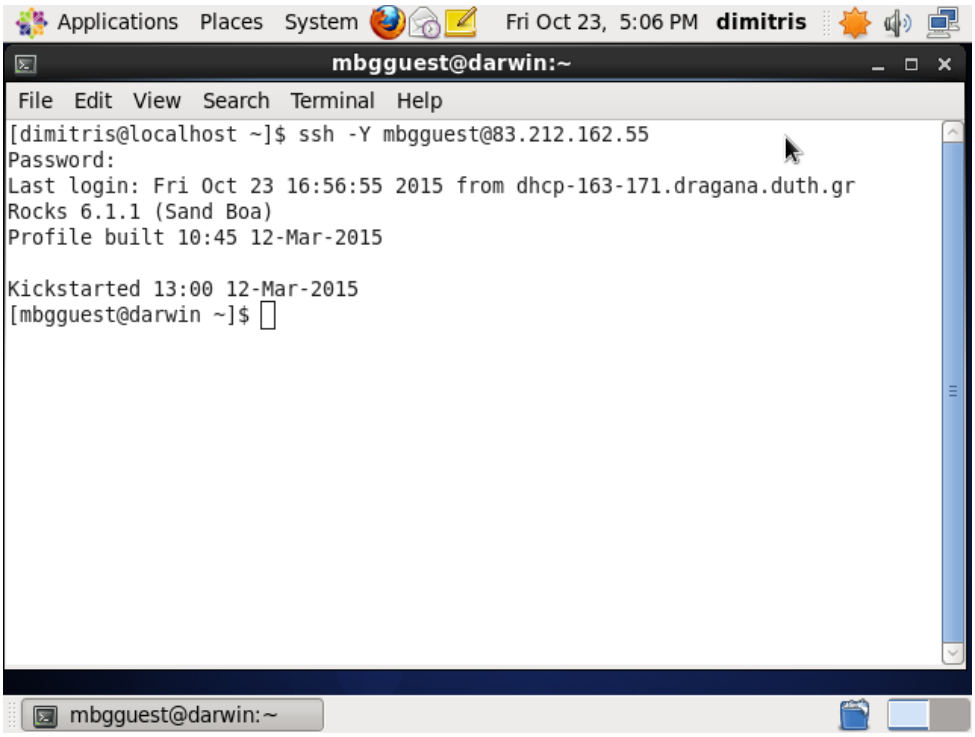
   Username: mbgguest **Password:** mbg2015guest

2. Applications → System Tools → Terminal
3. Στο παράθυρο που θα ανοίξει, πληκτρολογούμε:

   ssh -Y mbgguest@83.212.162.55

4. Password: mbg2015guest

   Το αποτέλεσμα παρουσιάζεται στην ακόλουθη οθόνη.



5. Στη συνέχεια πληκτρολογούμε:

   cd PaschouWS/user<u>Nr</u>

όπου **<u>Nr</u>** είναι ένας αριθμός που θα προσδιοριστεί κατά τη διάρκεια του workshop.

6. Για την επικύρωση του GWAS συνόλου δεδομένων και τη δημιουργία αρχείου με στατιστικά, πληκτρολογούμε:

PLINK 1.07

/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --out summary

PLINK 2.0

/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --out summary

7. Μελέτη των αποτελεσμάτων. Κάθε εκτέλεση του PLINK δημιουργεί ένα αρχείο καταγραφής (log file) που παρέχει χρήσιμες πληροφορίες για το σύνολο δεδομένων και την εντολή που εκτελέστηκε. Το αρχείο έχει την κατάληξη .log. Πιο σύνθετες εντολές του PLINK δημιουργούν επιπρόσθετα αρχεία που περιέχουν αποτελέσματα. Τα αρχεία αυτά έχουν διαφορετικές καταλήξεις ανάλογα με την κάθε εντολή.

8. Για το άνοιγμα του αρχείου με κατάληξη .log, πληκτρολογούμε:

   less summary.log

9. Εδώ παρουσιάζεται ένα πλήθος πληροφοριών σχετικά με το σύνολο δεδομένων GWASdata που επεξεργαζόμαστε.

***Missing Rates:* 0 of the ... individuals removed for low genotyping (MIND > 1)**

Το αρχικό βήμα στην ανάλυση όλων των δεδομένων είναι η εξαίρεση των ατόμων με πολύ μεγάλη απώλεια γονοτυπικών δεδομένων. Το MIND είναι το **M**aximum **IND**ividual missingness rate, π.χ. Η απώλεια δεδομένων SNP για ένα άτομο. Αρχικά τίθεται στην τιμή 1, επομένως, εξαιρούνται δείγματα με 100% απώλεια SNPs.

Συνήθως, χρησιμοποιούνται πιο αυστηροί έλεγχοι για τα προς επεξεργασία δεδομένα, όπως MIND > 0.1 Αυτό σημαίνει ότι θα αφαιρεθεί κάθε δείγμα που έχει απώλεια γονοτυπημένων SNPs μεγαλύτερη από 10%.

***Missing Rates:* 0 SNPs failed missingness test (GENO > 1)**

Το δεύτερο βήμα της ανάλυσης είναι η αφαίρεση SNPs που λείπουν από πολλά άτομα, για παράδειγμα εάν για ένα SNP οι τιμές λείπουν πάνω από το 10% των δειγμάτων. Εδώ ο έλεγχος γίνεται ανά στήλη και όχι ανά γραμμή. Αρχικά η τιμή GENO είναι μεγαλύτερη από 1, επομένως καμία στήλη SNP δεν αφαιρέθηκε.

Συνήθως, χρησιμοποιούνται πιο αυστηρά κριτήρια για τα προς επεξεργασία δεδομένα, όπως GENO > 0.1.

***Allele frequencies:* 0 SNPs failed frequency test (MAF <0)**

Το ακρωνύμιο MAF προέρχεται από τις λέξεις **M**inor **A**llele **F**requency. Το συγκεκριμένο στατιστικό μέγεθος χρησιμοποιείται για τον αποκλεισμό SNPs που δεν παρέχουν ικανοποιητική πληροφόρηση επειδή παρουσιάζουν μικρή διακύμανση στο προς ανάλυση σύνολο δεδομένων. Για παράδειγμα, αν ένα SNP παρουσιάζει παραλλαγή σε μόνο ένα άτομο του συνόλου μελέτης, δεν είναι στατιστικά σημαντικό και μπορεί να αφαιρεθεί από τα δεδομένα.

Αρχικά, η τιμή του MAF τίθεται μικρότερη από το μηδέν, γεγονός που σημαίνει ότι κανένα SNP δεν αποκλείεται. Συνήθως, χρησιμοποιούνται πιο αυστηροί έλεγχοι για τα προς επεξεργασία δεδομένα.

10. Για το κλείσιμο του αρχείου που άνοιξε με το *less summary.log* στο βήμα 8, πληκτρολογούμε το συνδυασμό: **CTRL+C**

11. Η λήψη στατιστικών στοιχείων αναφορικά με απώλειες στο σύνολο δεδομένων, μπορεί να γίνει με την εντολή:

PLINK 1.07

/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --missing --out summary_missing

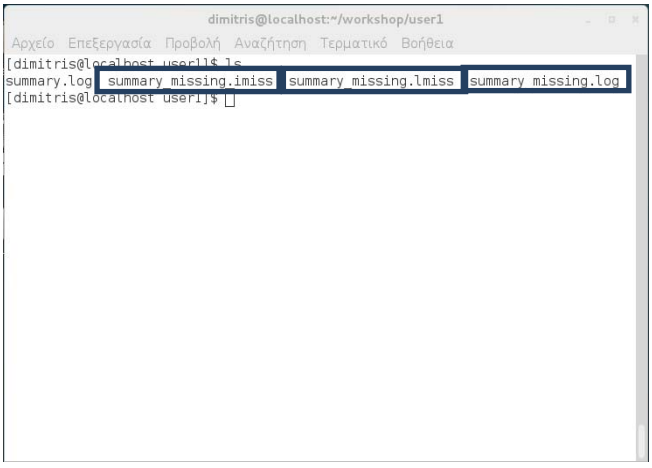
PLINK 2.0

/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --missing --out summary_missing

12. Τα ανά άτομο και ανά SNP ποσοστά που προκύπτουν, εντοπίζονται στα αρχεία **summary_missing.imiss** και **summary_missing.lmiss**, αντίστοιχα. Αυτά τα αρχεία είναι απλά αρχεία κειμένου που μπορούν να προβληθούν με έναν απλό κειμενογράφο.

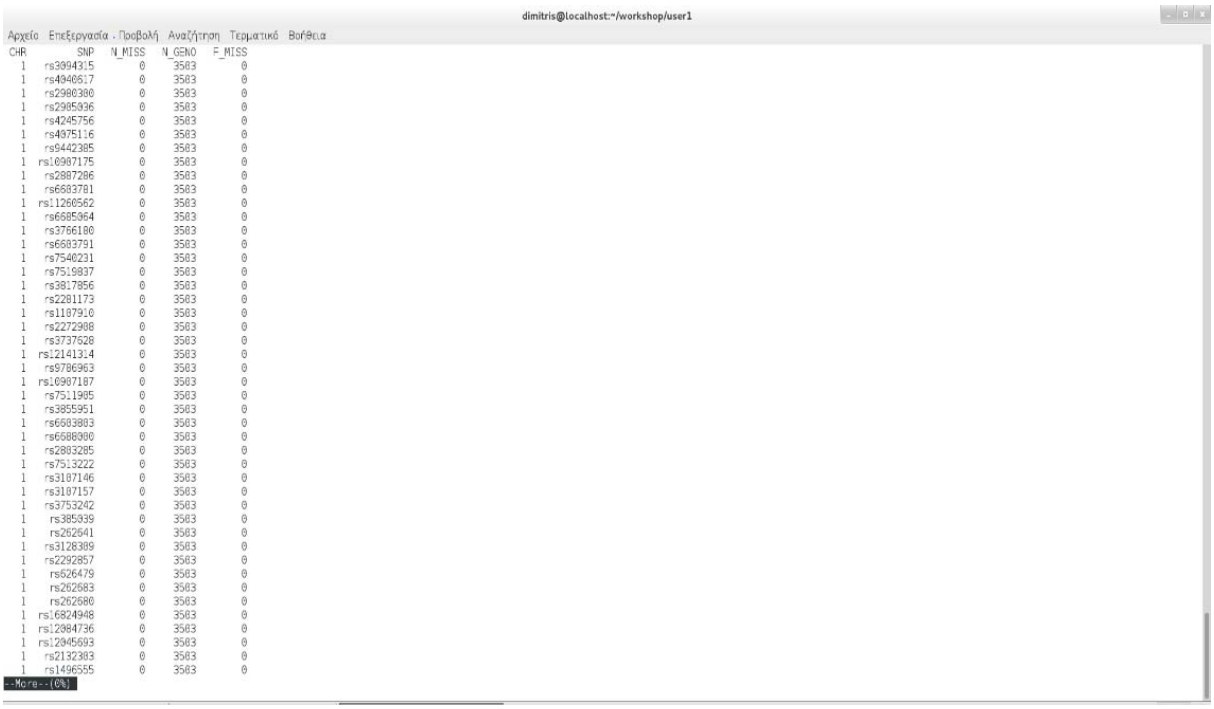
13. Για την προβολή των αρχείων στην οθόνη πληκτρολογούμε:
    ls

    Το αποτέλεσμα παρουσιάζεται στην ακόλουθη εικόνα.



14. Για την προβολή του αρχείου **summary_missing.lmiss** πληκτρολογούμε:
more summary_missing.lmiss

Τα αποτελέσματα που παρουσιάζονται στην οθόνη, δείχνουν για κάθε SNP, τον αριθμό (Number, N_MISS) και την αναλογία (Frequency, F_MISS) των εκλιπόντων ατόμων.



15. Για το κλείσιμο του αρχείου που άνοιξε, πληκτρολογούμε το συνδυασμό: **CTRL+C**

16. Για την προβολή του αρχείου **summary_missing.imiss** πληκτρολογούμε:
more summary_missing.imiss
Η τελευταία στήλη των αποτελεσμάτων που παρουσιάζονται στην οθόνη, αναφέρεται στο πραγματικό γονοτυπικό ποσοστό για κάθε άτομο.

17. Μπορεί να πραγματοποιηθεί ανάλυση των δεδομένων ανά χρωμόσωμα. Ενδεικτικά, στην περίπτωση ανάλυσης των δεδομένων αποκλειστικά για το χρωμόσωμα 1, τότε η εντολή που θα χρησιμοποιηθεί είναι:

PLINK 1.07

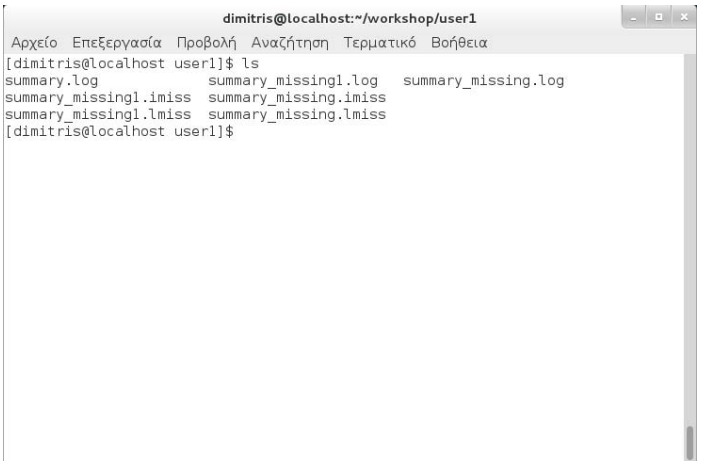
/home/mbgguest/PaschouWS/ Software/plink107/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --chr 1 --missing --out summary_missing1

PLINK 2.0

/home/mbgguest/PaschouWS/ Software/plink2/plink --noweb --bfile /home/mbgguest/PaschouWS/dataset/GWASdata --chr 1 --missing --out summary_missing1

18. Για την προβολή των αρχείων στην οθόνη πληκτρολογούμε:
    ls

    Το αποτέλεσμα παρουσιάζεται στην ακόλουθη εικόνα.



19. Πληκτρολογώντας:
    more summary_missing1.lmiss
    more summary_missing1.imiss

    Παρουσιάζονται τα αποτελέσματα για κάθε SNP στο χρωμόσωμα 1, και για το πραγματικό γονοτυπικό ποσοστό για κάθε άτομο για το χρωμόσωμα 1, αντίστοιχα.

20. Για το κλείσιμο του αρχείου που άνοιξε, πληκτρολογούμε το συνδυασμό: **CTRL+C.**

21. Ποιοτικός Έλεγχος του Dataset

Στα δεδομένα της GWAS μελέτης χρειάζεται να εφαρμοστεί πολλαπλός ποιοτικός έλεγχος για τη σωστή διεξαγωγή των ακολουθούμενων στατιστικών μεθόδων και την αποφυγή ψευδώς θετικών συσχετίσεων.

Τα βασικά βήματα στον ποιοτικό έλεγχο (quality control – QC) είναι τα εξής:

- Minor Allele Frequency (--maf)
- Genotyping Missingness (--geno)
- Individual Missingness (--mind)
- Hardy Weinberg (--hwe)

Για την εφαρμογή κάθε μιας από τις προαναφερθείσες παραμέτρους ποιοτικού ελέγχου, χρησιμοποιείται μια εντολή με την ακόλουθη σύνταξη:

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata {παράμετρος}
τιμή_παραμέτρου --out test_{όνομα_παραμέτρου}
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata {παράμετρος}
τιμή_παραμέτρου --make-bed --out test_{όνομα_παραμέτρου}
```

Όπου {**παράμετρος**}: --maf, --geno, --mind και --hwe

τιμή_παραμέτρου είναι η τιμή που ορίζεται για τη συγκεκριμένη παράμετρο.

Παρατήρηση: Όταν χρησιμοποιείται η έκδοση **PLINK 2.0**, τότε είναι απαραίτητη η παράμετρος **--make-bed** για τη δημιουργία των αρχείων bed, bim και fam. Στην έκδοση **PLINK 1.07 δεν** χρειάζεται η συγκεκριμένη παράμετρος (--make-bed) και το αποτέλεσμα που προκύπτει είναι ένα αρχείο καταγραφής (log file) με τα αποτελέσματα.

Δοκιμάζοντας κάθε μια παράμετρο ξεχωριστά ώστε να προκύψουν τα αντίστοιχα αρχεία και να γίνει αντιληπτή η επίπτωσή της στα διαθέσιμα δεδομένα.

Minor Allele Frequency (--maf)

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --maf 0.01 --out test_maf
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --maf 0.01 --make-bed --out
test_maf
```

Genotyping Missingness (--geno)

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --geno 0.02 --out test_geno
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --geno 0.02 --make-bed --out
test_geno
```

Individual Missingness (--mind)

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --mind 0.02 --out test_mind
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --mind 0.02 --make-bed --out
test_mind
```

Hardy Weinberg (--hwe)

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --hwe 0.001 --out test_hwe
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --hwe 0.001 --make-bed --out
test_hwe
```

Οι μεμονωμένοι, προηγούμενοι ποιοτικοί έλεγχοι μπορούν να εκτελεστούν ταυτόχρονα σε μια εντολή και να προκύψουν συγκεντρωτικά αποτελέσματα ή/και συγκεντρωτική ομάδα αρχείων bed, bim, fam.

PLINK 1.07 Χωρίς δημιουργία των αρχείων bed, bim, fam

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --maf 0.01 --geno 0.02 --mind
0.02 --hwe 0.001 --out GWASdata_qc
```

PLINK 1.07 Με δημιουργία των αρχείων bed, bim, fam

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --maf 0.01 --geno 0.02 --mind
0.02 --hwe 0.001 --make-bed --out GWASdata_qc
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
/home/mbgguest/PaschouWS/dataset/GWASdata --maf 0.01 --geno 0.02 --mind
0.02 --hwe 0.001 --make-bed --out GWASdata_qc
```

22. Ανάλυση συσχέτισης ενός SNP

Για την εκτέλεση μιας βασικής αλληλικής δοκιμασίας συσχέτισης για ένα SNP τη φορά με το φαινότυπο, πληκτρολογούμε την ακόλουθη εντολή:

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
GWASdata_qc --assoc --adjust --out assoc1
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
GWASdata_qc --assoc --adjust --out assoc1
```

Η χρήση της παραμέτρου --adjust δίνει εντολή στο plink να διορθώσει τις τιμές p (p-values) που προκύπτουν από τη δοκιμασία συσχέτισης. Επίσης, αναφέρει και τον παράγοντα πληθωρισμού (lambda) στο αρχείο log. Οι τιμές p μπορούν να απεικονιστούν είτε στο Haploview, είτε χρησιμοποιώντας κάποιο script (π.χ. Python, R).

Το αρχείο .assoc περιέχει τις τιμές τις δοκιμασίας συσχέτισης για κάθε SNP. Η δομή του μοιάζει με την ακόλουθη:

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3094315	792429	G	0.1489	0.08537	A	1.684	0.1944	1.875
1	rs4040617	819185	G	0.1354	0.08537	A	1.111	0.2919	1.678
1	rs4075116	1043552	C	0.04167	0.07317	T	0.8278	0.3629	0.5507
1	rs9442385	1137258	T	0.3723	0.4268	G	0.5428	0.4613	0.7966
1	rs11260562	1205233	A	0.02174	0.03659	G	0.3424	0.5585	0.5852
1	rs6685064	1251215	C	0.3854	0.439	T	0.5253	0.4686	0.8013
1	rs3766180	1563420	T	0.1771	0.09756	C	2.317	0.128	1.991

Το CHR δηλώνει το χρωμόσωμα και το BP τη θέση του SNP πάνω στο χρωμόσωμα. Το A1 δηλώνει το Minor Allele και το A2 το Major Allele. F_A και F_U είναι οι συχνότητες του A1 στους ασθενείς και στους μάρτυρες αντίστοιχα. CHISQ, P, OR είναι τα αποτελέσματα της δοκιμασίας.

Για την απόκτηση ενός πιο εύκολου στην ανάγνωση αρχείου, μπορεί να προστεθεί και η παράμετρος:

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
GWASdata_qc --assoc --adjust --pfilter 1e-5 --out assoc1_filtered
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
GWASdata_qc --assoc --adjust --pfilter 1e-5 --out assoc1_filtered
```

Το αποτέλεσμα από την προσθήκη της παραμέτρου **--pfilter 1e-5**, είναι το τελικό αρχείο να περιέχει μόνο τα SNPs που έχουν p-value μικρότερη από 10^{-5}.

Ο παράγοντας πληθωρισμού συγκρίνει όλες τις τιμές p που αποκτήθηκαν από τη δοκιμασία με βάση το τι περιμέναμε να αποκτήσουμε. Αν η τιμή του παράγοντα είναι αρκετά μεγαλύτερη από το 1.00 (>1.1), τότε σημαίνει πως είχαμε περισσότερες απ' ότι θα περιμέναμε και πρέπει να το διορθώσουμε.

Τις περισσότερες φορές, ο πληθωρισμός στις τιμές p προκαλείται από την παρουσία διαφορετικών πληθυσμών στο ίδιο σετ δεδομένων. Τα επόμενα βήματα οδηγούν στη διόρθωση αυτού του προβλήματος.

23. Διόρθωση με χρήση λογαριθμικής παλινδρόμησης

Η μέθοδος της λογαριθμικής παλινδρόμησης επιτρέπει τη χρήση επιπρόσθετων συμμεταβλητών όταν εκτελεστεί η δοκιμασία συσχέτισης ενός SNP με το φαινότυπο. Οι συμμεταβλητές μπορούν να είναι είτε συνεχείς είτε διακριτές. Αυτή η μέθοδος είναι πιο ευέλικτη, αλλά τρέχει πιο αργά από τη βασική εντολή –assoc.

Η λογαριθμική παλινδρόμηση εκτελείται με την εντολή:

PLINK 1.07

/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile GWASdata_qc --logistic --covar {το αρχείο των συμμεταβλητών} --hide-covar --out assoc2

PLINK 2.0

/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile GWASdata_qc --logistic --covar {το αρχείο των συμμεταβλητών} --hide-covar --out assoc2

Οι συμμεταβλητές που θα ενσωματώσουμε παράγονται από το EIGENSOFT. Για να τρέξει το πρόγραμμα EIGENSOFT, το μόνο που χρειάζεται είναι μια τριάδα αρχείων γονοτυπικών δεδομένων (bed, bim, fam) και ένα αρχείο parfile που δίνει τις εντολές. Το αρχείο **parfile** έχει τη δομή:

genotypename: GWASdata_pruned.bed
snpname: GWASdata_pruned.bim
indivname: GWASdata_pruned.fam
evecoutname: output_pca.evec
evaloutname: output_pca.eval
familynames: YES

Τα επόμενα βήματα δείχνουν τη διαδικασία παραγωγής του αρχείου .parfile και την εκτέλεση λογαριθμικής παλινδρόμησης με την εντολή PLINK

24. Για τη δημιουργία της τριάδας γονοτυπικών δεδομένων, θα χρειαστεί να αραιώσουμε το σετ. Η αραίωση γίνεται με έλεγχο των SNPs που βρίσκονται σε ανισορροπία σύνδεσης (Linkage Disequilibrium - LD pruning).

Η παράμετρος **--indep** μειώνει (prune) τα δεδομένα βάση του συντελεστή διακύμανσης πληθωρισμού ή συντελεστή διογκώσεως της διακυμάνσεως (**V**ariance **I**nflation **F**actor - VIF). Ο VIF εκφράζει το ρυθμό με τον οποίο αυξάνεται η διακύμανση ενός εκτιμητή όταν υπάρχει πολυσυγγραμικότητα. Το VIF είναι ίσο με:

$VIF = 1/(1 - R^2)$

Όπου R^2 είναι ο πολλαπλός συντελεστής συσχέτισης για ένα SNP σε σχέση με όλα τα υπόλοιπα SNPs.

Η παράμετρος **--indep** ακολουθείται από τρεις αριθμούς, ως εξής:

--indep n1 n2 n3

όπου

n1: Το παράθυρο των SNPs, δηλαδή τα SNPs που θα εξετάζονται σε κάθε βήμα.

n2: Ο αριθμός των SNPs κατά τον οποίον μετατοπίζεται το παράθυρο σε κάθε βήμα.

n3: Το κατώφλι του VIF.

Επίσης, υπάρχει η παράμετρος **--indep-pairwise** η οποία είναι παρόμοια με την προηγούμενη, όμως, υπάρχει η διαφοροποίηση ότι βασίζεται μόνο σε γονοτυπική συσχέτιση ζευγών.

Η παράμετρος **--indep-pairwise** ακολουθείται από τρεις αριθμούς, ως εξής:

--indep n1 n2 n3

όπου

n1: Το παράθυρο των SNPs, δηλαδή τα SNPs που θα εξετάζονται σε κάθε βήμα.

n2: Ο αριθμός των SNPs κατά τον οποίον μετατοπίζεται το παράθυρο σε κάθε βήμα.

n3: είναι το κατώφλι r^2. Το μέγεθος αυτό αποτελεί το μέτρο σύγκρισης μεταξύ δυο SNPs, ενώ δεν είναι ο πολλαπλός συντελεστής συσχέτισης.

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile
GWASdata_qc --indep-pairwise 50 5 0.1 --out pruning
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile
GWASdata_qc --indep-pairwise 50 5 0.1 --out pruning
```

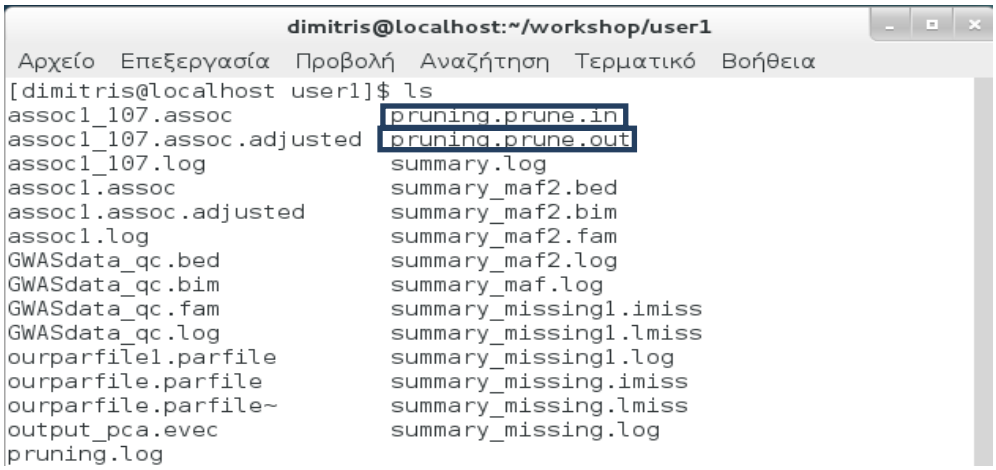
Η εκτέλεση της προηγούμενης εντολής έχει σαν αποτέλεσμα τη δημιουργία δυο αρχείων, με καταλήξεις **.prune.in** και **.prune.out**.

Το αρχείο με κατάληξη **.prune.in** περιέχει τα SNPs που **δεν** βρίσκονται σε ανισορροπία σύνδεσης. Συνεπώς, είναι τα SNPs που χρειάζονται για τη συνέχιση της μελέτης.

Το αρχείο με κατάληξη **.prune.out** περιέχει τα SNPs που βρίσκονται σε ανισορροπία σύνδεσης. Επομένως, είναι τα SNPs που μπορούν να παραληφθούν από το σύνολο δεδομένων, καθώς φαίνεται να μην περιέχουν χρήσιμη πληροφορία για τη μελέτη.

25. Για την προβολή των σχετικών αρχείων στην οθόνη πληκτρολογούμε:
    ls

Το αποτέλεσμα είναι να υπάρχουν δυο αρχεία που είναι: **pruning.prune.in** και **pruning.prune.out**. Το pruning είναι το όνομα του αρχείου, που ορίζεται με την παράμετρο --out



26. Για τη δημιουργία του συνόλου δεδομένων με τα SNPs που δεν είναι σε ανισορροπία σύνδεσης, πληκτρολογούμε την ακόλουθη εντολή

PLINK 1.07

/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile GWASdata_qc --extract pruning.prune.in --make-bed --out GWASdata_pruned

PLINK 2.0

/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile GWASdata_qc --extract pruning.prune.in --make-bed --out GWASdata_pruned

27. Δημιουργούμε το αρχείο **ourparfile.parfile** πληκτρολογώντας (ένα εκ των δυο εντολών)

gedit	**nano**
Ανοίγει ο κειμενογράφος, όπου πληκτρολογείτε το ακόλουθο κείμενο:	
genotypename: GWASdata_pruned.bed snpname: GWASdata_pruned.bim indivname: GWASdata_pruned.fam evecoutname: output_pca.evec evaloutname: output_pca.eval familynames: YES	
Επιλέγουμε το εικονίδιο: **Αποθήκευση** Δίνουμε το όνομα: ourparfile.parfile Αποθηκεύουμε το αρχείο στο φάκελο που δουλεύουμε (userNR) Κλείνουμε το πρόγραμμα	**CTRL+O** για αποθήκευση Δίνουμε το όνομα: /home/mbgguest/PaschouWP/userNR /ourparfile.parfile Πατάμε το πλήκτρο ENTER **CTRL+X** για έξοδο

28. Εκτέλεση του προγράμματος EIGENSOFT ώστε να γίνει υπολογισμός των συμμεταβλητών που θα χρησιμοποιηθούν στην εφαρμογή της λογαριθμικής παλινδρόμησης.

EIGENSOFT

/home/mbgguest/PaschouWS/Software/ EIG6.0.1/bin/smartpca –p ourparfile.parfile

29. Χρησιμοποιώντας την εντολή **AWK** το αρχείο αποκτά την κατάλληλη μορφή που απαιτείται από το plink.

awk

awk –F":" '{print $1, $2}' output_pca.evec | tail –n +2 | awk '{print $1, $2, $3, $4, $5, $6, $7}' > output_pca.covar

30. Τώρα μπορούμε να εκτελέσουμε τη λογαριθμική παλινδρόμηση σύμφωνα με την περιγραφή στο βήμα 23.

PLINK 1.07

/home/mbgguest/PaschouWS/Software/plink107/plink --noweb --bfile GWASdata_qc --logistic --covar output_pca.covar --hide-covar --out assoc2

PLINK 2.0

/home/mbgguest/PaschouWS/Software/plink2/plink --noweb --bfile GWASdata_qc --logistic --covar output_pca.covar --hide-covar --out assoc2

31. Σύνοψη αποτελεσμάτων

Αντί να αναφέρουμε μια μεγάλη λίστα από SNPs που έχουν συσχετιστεί με το φαινότυπο, από τα οποία η πλειοψηφία βρίσκεται σε ανισορροπία σύνδεσης, μερικές φορές είναι πρακτικό να συνοψίζεται το αποτέλεσμα των δοκιμασιών συσχέτισης, ομαδοποιώντας τα SNPs σε συσσωματώματα. Μπορούμε επίσης να παρέχουμε και τις γονιδιακές συντεταγμένες ώστε να πάρουμε πληροφορίες για τα γονίδια που βρίσκονται στην περιοχή.

32. Από τον ιστότοπο του plink (**http://pngu.mgh.harvard.edu/~purcell/plink/dist/glist-hg18**) επιλέγουμε όλα τα στοιχεία με **CTRL+A** και στη συνέχεια τα αντιγράφουμε με **CTRL+C**

33. Δημιουργούμε το αρχείο πληκτρολογώντας (ένα εκ των δυο εντολών)

gedit	**nano**
Ανοίγει ο κειμενογράφος	
CTRL+V	
Επιλέγουμε το εικονίδιο: **Αποθήκευση** Δίνουμε το όνομα: **glish-hg18.txt** Αποθηκεύουμε το αρχείο στο φάκελο που δουλεύουμε (userNR) Κλείνουμε το πρόγραμμα	**CTRL+O** για αποθήκευση Δίνουμε το όνομα: **/home/mbgguest/PaschouWP/userNR/ glish-hg18.txt** Πατάμε το πλήκτρο ENTER **CTRL+X** για έξοδο

34. Εκτελούμε την ακόλουθη εντολή:

PLINK 1.07

```
/home/mbgguest/PaschouWS/Software /plink107/plink --noweb --bfile GWASdata_qc --clump assoc2.assoc.logistic --clump-p1 1e-5 --clump-kb 500 --clump-r2 0.05 –-clump-range glist-hg18.txt --out clumping
```

PLINK 2.0

```
/home/mbgguest/PaschouWS/Software /plink2/plink --noweb --bfile GWASdata_qc --clump assoc2.assoc.logistic --clump-p1 1e-5 --clump-kb 500 --clump-r2 0.05 –-clump-range glist-hg18.txt --out clumping
```

35. Για την προβολή του αρχείου **clumping.clumped** πληκτρολογούμε:
more clumping.clumped

Το αρχείο clumping.clumped που προκύπτει έχει τη δομή:

CHR	F	SNP	BP	P	TOTAL	NSIG	S05	S01	S001	S0001	SP2
8	1	rs1234564	15716326	5.01e-07	0	0	0	0	0	0	NONE
14	1	rs1205236	69831825	1.46e-06	0	0	0	0	0	0	NONE
2	1	rs16331058	114547107	2.33e-06	3	0	0	0	0	3	rs2366902(1),...
2	1	rs759966	54902416	9.28e-06	4	0	0	0	3	1	rs12538389(1),...
11	1	rs8031586	44633498	9.75e-06	1	0	0	0	0	1	rs802328(1)
12	1	rs12431413	30028246	9.89e-06	0	0	0	0	0	0	NONE
6	1	rs14966070	62091121	1.07e-05	0	0	0	0	0	0	NONE

36. Για το κλείσιμο του αρχείου που άνοιξε, πληκτρολογούμε το συνδυασμό: **CTRL+C**

37. Για την προβολή του αρχείου **clumping.clumped.ranges** πληκτρολογούμε:
more clumping.clumped.ranges

Το αρχείο clumping.clumped.ranges που προκύπτει έχει τη δομή:

CHR	SNP	P	N	POS	KB	RANGES
17	rs9944528	1.927e-05	2	chr17:77894039..77933018	38.979	[UTS2R,SKIP,FLJ35767]
9	rs17534370	1.958e-05	1	chr9:70297172..70297172	0	[PGM5]
11	rs12418173	1.965e-05	7	chr11:112102294..112133479	31.185	[]

38. Από το αρχείο **clumping.clumped.ranges** επιλέγονται τα γονίδια από τη δεξιά στήλη (χωρίς τις αγκύλες) και μεταφέρονται με αντιγραφή και επικόλληση, σε ένα νέο αρχείο, μέσα στο φάκελο userNR.

39. Για το κλείσιμο του αρχείου που άνοιξε, πληκτρολογούμε το συνδυασμό: **CTRL+C**

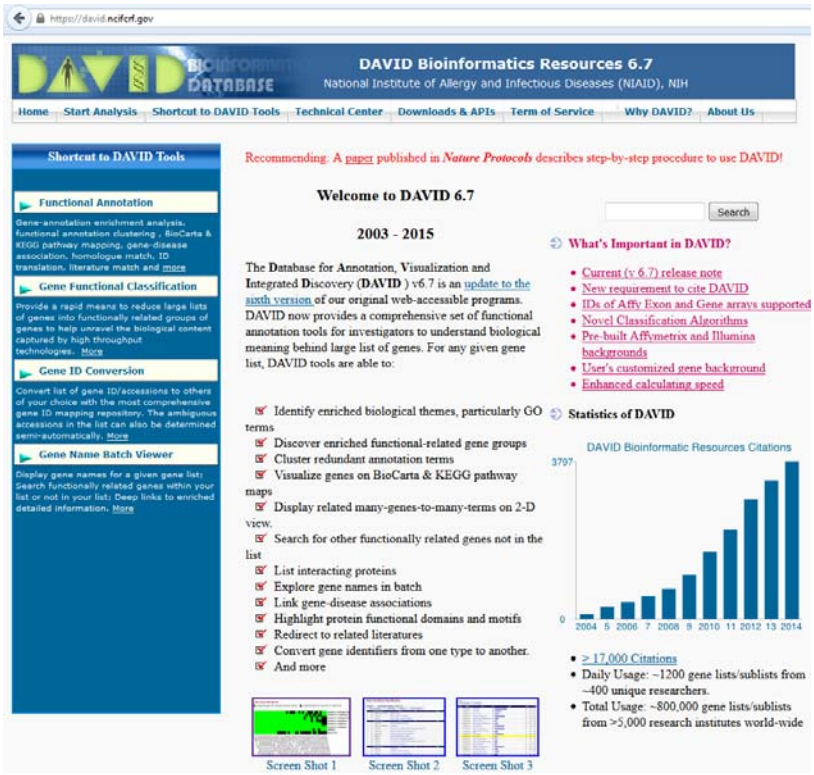
40. Από τα αποτελέσματα του clumping, μπορούμε να εξάγουμε τα γονίδια και να ελέγξουμε αν ανήκουν σε κάποια από τα γνωστά μονοπάτια κυτταρικών διεργασιών. Αυτό γίνεται πληκτρολογώντας:

awk '{print $7}' clumping.clumped.ranges | sed "s/[][]//g" | grep –v '^$' > genranges

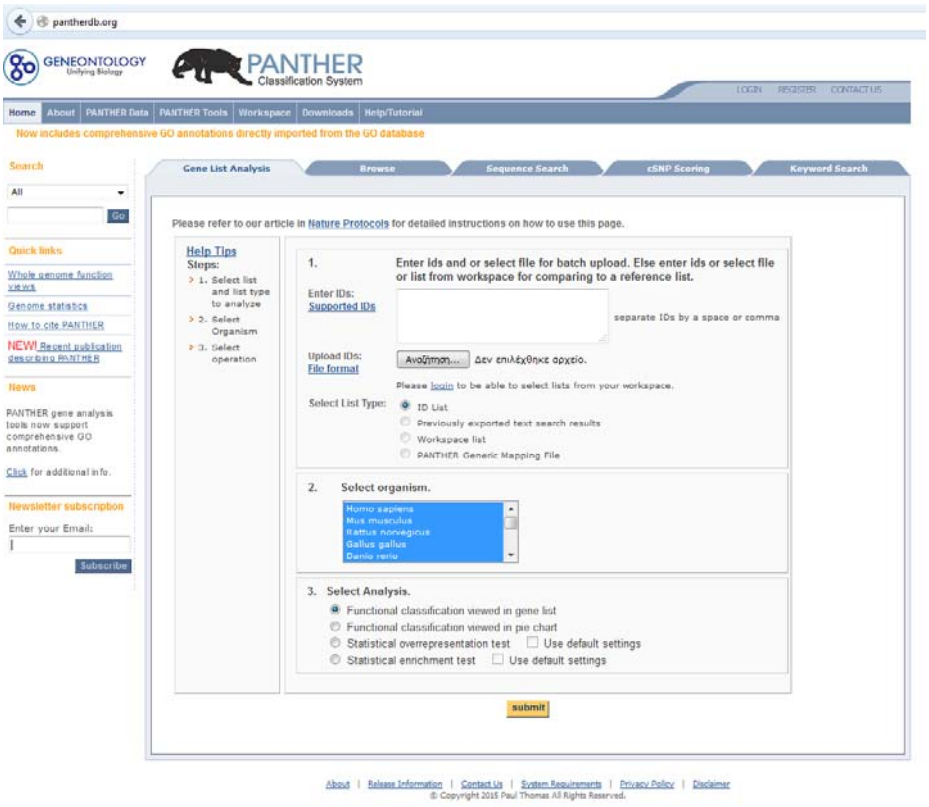
41. Ανάλυση μονοπατιών

Από τα αποτελέσματα του clumping, μπορούμε να εξάγουμε τα γονίδια και να ελέγξουμε αν ανήκουν σε κάποια από τα γνωστά μονοπάτια κυτταρικών διεργασιών. Σε αυτό θα μας βοηθήσουν τρία διαδικτυακά εργαλεία:

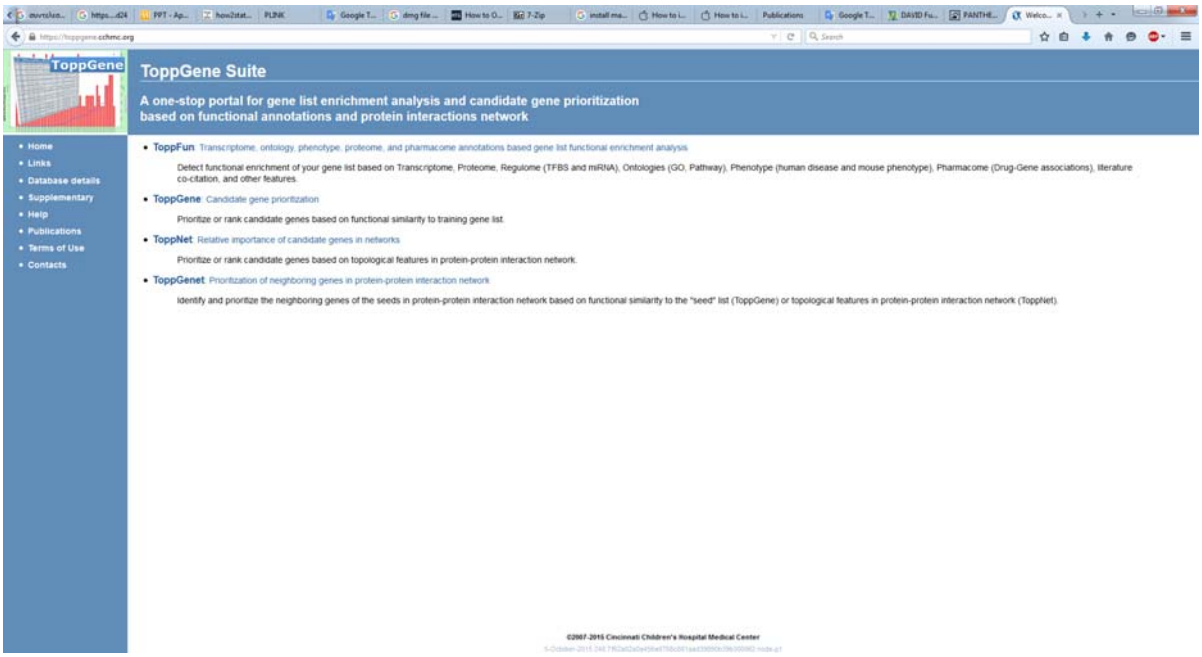
- DAVID: david.ncifcrf.gov
- PANTHER: pantherdb.org
- ToppGene: toppgene.cchmc.org



DAVID: david.ncifcrf.gov



PANTHER: pantherdb.org



ToppGene: toppgene.cchmc.org

Η ανάλυση μονοπατιών βοηθάει να εξετάσουμε τη βιολογική σημασία των αποτελεσμάτων μας, ώστε αυτά να γίνουν κατανοητά.