

Επεξεργασία των αρχείων

Στη διάθεση μας έχουμε 22 χρωμοσώματα, όπου κάθε χρωμόσωμα έχει ένα αρχείο .bed, .fam και .bim

● .bim:

Το αρχείο αυτό περιέχει 6 στήλες. Η πρώτη στήλη περιέχει τον αριθμό του χρωμοσώματος στο οποίο αναφέρεται το αρχείο. Η δεύτερη στήλη περιέχει το κώδικό του **snp**, η τρίτη στήλη περιέχει τη θέση του snp, η τέταρτη στήλη περιέχει το **base-pair**, η πέμπτη στήλη περιέχει το **Allele1** που είναι συνήθως η μειονότητα και η έκτη στήλη περιέχει το **Allele2** που είναι η πλειοψηφία. Περισσότερα μπορείτε να βρείτε στην σελίδα: <https://www.cog-genomics.org/plink2/formats#bim>

● .fam:

Το αρχείο αυτό περιέχει 6 στήλες. Η πρώτη στήλη έχει το **Family ID**, η δεύτερη στήλη περιέχει το **ID** του ασθενούς (στην περίπτωση μας επειδή όλοι οι ασθενείς δεν ανήκουν στη ίδια οικογένεια το **Family ID** είναι ίδιο με το **ID**). Η τρίτη στήλη το id του πατέρα και η τέταρτη στήλη το id της μητέρας. Η πέμπτη στήλη μας δίνει την πληροφορία για το φύλλο του ασθενή και η έκτη στήλη έχει πληροφορία για τον φενότυπο (στη δικιά μας περίπτωση πληροφορία για τον φενότυπο αντλούμε από άλλο αρχείο που θα δούμε παρακατάτω) Περισσότερα μπορείτε να βρείτε στην σελίδα <https://www.cog-genomics.org/plink2/formats#fam>

● .bed:

Το αρχείο **.bed** περιέχει όλη τη γενετική πληροφορία του χρωμοσώματος που είναι κωδικοποιημένη στη δεκαεξαδική μορφή. Περισσότερα μπορείτε να βρείτε στις σελίδες <https://www.cog-genomics.org/plink2/formats#bed> και <https://genome.ucsc.edu/FAQ/FAQformat#format1>

Πληροφορία για τους φενότυπους αντλούμε από το αρχείο **phenotype_euro_edit.txt** το οποίο έχει ασθενείς από την ευρώπη. Περιέχει 4980 ασθενείς από τους οποίους οι 1018 έχουν **case = 1** και οι υπόλοιποι 3962 έχουν **control = 0**. Αυτό το αρχείο έχει την εξής μορφή:

- στη πρώτη στήλη έχει το **eid** που είναι το id του ασθενή
- στη δεύτερη στήλη έχει το φύλλο του ασθενή (0 = γυναίκα, 1 = άντρας)
- στη τρίτη στήλη έχει το έτος που γεννήθηκε
- και στη τρίτη στήλη έχει το φενότυπο (0 = control, 1 = case)

Για να κάνουμε την επεξεργασία των αρχείων **.bed**, **.fam** και **.bim**, θα χρησιμοποιήσουμε το πρόγραμμα **plink**. Μπορείτε να το βρείτε εδώ: <https://www.cog-genomics.org/plink2>.

Αρχικά τρέξαμε την εντολή **plink --bfile chrX --allow-no-sex --out chrX --1 --pheno phenotype_edit.txt --assoc**

- **--bfile chrX** με αυτή την εντολή διαλέγεις ποιο χρωμόσωμα θες να επεξεργαστείς
- **--allow-no-sex** αυτή την επιλογή την βάζουμε για να μην λάβει υπόψη το φύλλο
- **--out chrX** σε ποιο αρχείο θα αποθηκευτούν τα δεδομένα της επεξεργασίας
- **--1** επειδή οι φενότυποι που έχουμε είναι κωδικοποιημένοι σε 0 και 1 βάζουμε αυτήν την παράμετρο για να την καταλάβει τη plink
- **--pheno phenotype_edit.txt** το αρχείο όπου διαβάσει το φενότυπο. Προσοχή πρέπει να είναι της μορφής (**FID, IID, CASE/CONTROL**) όπου στη περίπτωση μας το **FID** είναι ίδιο με το **IID**.
- **--assoc** είναι ο τύπος του πειράματος που τρέχουμε

Το **assoc** αρχείο είναι της μορφής:

CHR	Chromosome code
SNP	Variant identifier
BP	Base-pair coordinate
A1	Allele 1 (usually minor)
F_A	Allele 1 frequency among cases
F_U	Allele 1 frequency among controls
A2	Allele 2
CHISQ	Allelic test chi-square statistic. <i>Not present with 'fisher'/'fisher-midp' modifier.</i>
P	Allelic test p-value
OR	odds(allele 1 case) / odds(allele 1 control)

Περισσότερα για το **assoc** μπορείτε να βρείτε εδώ: <https://www.cog-genomics.org/plink/1.9/formats#assoc>

Επίσης τρέξαμε την εντολή **plink --recode lgen --out chrX --keep-fam patient.txt --bfile chrX --1 --allow-no-sex --extract chrXsnplist.txt** για να πάρουμε τα **snps** του κάθε ασθενή με τα **alleles** τους για το κάθε χρωμόσωμα.

- **--bfile chrX** με αυτή την εντολή διαλέγεις ποιο χρωμόσωμα θες να επεξεργαστείς
- **--allow-no-sex** αυτή την επιλογή την βάζουμε για να μην λάβει υπόψη το φύλλο
- **--out chrX** σε ποιο αρχείο θα αποθηκευτούν τα δεδομένα της επεξεργασίας
- **--1** επειδή οι φενότυποι που έχουμε είναι κωδικοποιημένοι σε 0 και 1 βάζουμε αυτήν την παράμετρο για να την καταλάβει τη plink
- **--recode lgen** είναι η παραμέτρος για να εξάγουμε το αρχείο που θα έχει τους ασθενείς με τα snps και τα alleles τους για το χρωμόσωμα X
- **--keep-fam patient.txt** με αυτήν την παράμετρο εξετάζουμε τους ασθενείς που βρίσκονται στο αρχείο **patient.txt**
- **--extract chrXsnplist.txt** με αυτήν τη παράμετρο εξετάζουμε τα snps που βρίσκονται στο αρχείο **chrXsnplist.txt**

Από αυτήν την εντολή παράγεται ένα αρχείο **.lgen** το οποίο είναι της μορφής:

1. **Family ID**
2. **Within-family ID**
3. **Variant identifier**
4. **Allele 1**
5. **Allele 2**

Περισσότερα για το **lgen** μπορείτε να βρείτε εδώ: <https://www.cog-genomics.org/plink/1.9/formats#lgen>

Το snp κάθε ασθéné παίρνει τον κωδικό 0 αν το allele1 **και** το allele2 απο το lgen αρχείο είναι διαφορετικο από το allele1(A1) του assoc αρχείου, παίρνει τον κωδικό 1 αν **μόνο** ένα από τα allele1 και allele2 απο το lgen αρχείο είναι ίδια με το allele1(A1) του assoc αρχείου και τέλος παίρνει τον κωδικό 2 αν το allele1 **και** το allele2 απο το lgen αρχείο είναι ίδια με το allele1(A1) του assoc αρχείου. Τον τρόπο για να τα κωδικοποιήσουμε έτσι το βρήκαμε στο παρακάτω αρχείο: <http://www.diva-portal.org/smash/get/diva2:845171/FULLTEXT01.pdf> στο κεφάλαιο 3.1 .