Homework 2

Introduction

We simulated data on a number of families, including parents and/or a number of children. The data file pedigree.ped includes relevant data about the family structure, genetic markers and a continuous phenotype. A second data file, pedigree.map, contains data about the genetic markers. These two file formats ped and map are standard file formats to work with PLINK but can also be used for analysis in R.

Specific questions

1. Look up what the general structure of a ped file, e.g. pedigree.ped, is. How is missing data indicated in data? What is the missing data indicator standardly used by the R software?
2. How is the data set "pedigree.ped" composed? Use R here to get to know your data.
   - 2.1. Describe the family structure, e.g., how many families, how many parents (couples of father and mother), how many children, how many children on average per family?
   - 2.2. What is the distribution of males and females in the data set?
   - 2.3. How many genetic markers have been generated? What are the minor allele frequencies of SNP1, SNP2, SNP3, SNP4 and SNP5? Is there a difference in allele frequency when the entire data are used, or when only the parents are used? What do you observe?
   - 2.4. How is the continuous phenotype distributed?
3. What are the inbreeding coefficients for the data at hand (use for example pedigree package in R)? What do the inbreeding coefficients tell you?
4. Perform a quality control on the data for a subsequent analysis (population based genetic association testing). Make a summary of the quality control. Use Plink or R tools.
   - 4.1. Remove all SNPs with a call rate < 0.95, i.e., a missingness of more than 5% (if applicable).
   - 4.2. Remove all individuals with a call rate < 0.95, i.e., a missingness of more than 5% (if all applicable).
   - 4.3. Remove all individuals with minor allele frequency lower than 0.01.
   - 4.4. Remove all genetic markers which are out of Hardy-Weinberg equilibrium (use a p-value of 0.0001).
5. Perform a genetic association analysis to identity important genetic markers associated with the continuous phenotype. Which part of the data can be used for this analysis? Use Plink or R tools. What are the results?

Write a small report, including some explanations about how you obtained the results. Send your answers to bmaus@ulg.ac.be.

Due date: 15 April 2013