Homework 1

Introduction

We simulated data on a number of families, including parents and/or a number of children. The data file pedigree2.ped includes relevant data in "LINKAGE" format, and hence also data on a number of genetic markers. A second data file, pheno.dat, includes data on a continuous phenotype. These data can be linked to the pedigree data using the unique identifiers for family members.

Specific questions

Q1. Look up what it means "LINKAGE" format. After having consulted some guidelines on LINKAGE formats, do you understand the structure of the pedigree file? Also, how is missing data indicated in the data? Are there different missing data indicators for the first 6 columns than for the marker data? What is the missing data indicator standardly used by the R software?

Q2. How many families have been generated?

Q3. What is the average number of kids per family?

Q4. Are there parents without children? If so, how many? How many parents (couples of father and mother) are there in total in the data set?

Q5. What is the distribution of males and females in the total data set? What is the distribution of males and females among the children? What is the indicator used for "males" / "females"?

Q6. How many markers have been generated?

Q7. What is the minor allele frequency of SNP1, of SNP2, SNP3, SNP4 and SNP5? Can you find a function to automate this computation? Is there a difference in allele frequency when the entire data are used, or whether only the parents are used? What do you observe?

Write a small report, including some explanations about how you obtained the answers

Due date: 25 February 2011