Homework 4A

Data analysis project

Introduction

We simulated data on a number of families, including parents and/or a number of children. The data file pedigree.dat includes relevant data in "LINKAGE" format (look up what this format means …), 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.

In this homework it is the idea to apply the R "SNPassoc" package to simulated data, while answering the questions below. However, if you can answer the questions using another technique you have seen in other classes, this is fine as well.

The first important hurdle to overcome, is to transform the data into a format that can be handled by SNPassoc. This can be quite time consuming. Do not be disappointed… More than half of the analysis time of a bioinformatician involves data manipulation!

Finally note that this homework 4A restricts attention to parents only. Hence make a subselection of your data to parents only. Homework 4B will use all of the data.

Specific questions on population-based genetic association analysis

Q1. Perform a data quality control check:

- Do a genome-wide check on the validity of the HWE condition. Which markers do not satisfy this condition? [Note: Use a function which is able to check this condition using one line only]
- Are there genotypes missing? Do you observe particular patterns in the observed/missing genotypes?
- Are there monomorphic SNPs? How should they be treated further on in subsequent analyses?

Q2 In the phenotype-file, a continuous trait has been provided. Perform a genomewide scan for this continuous trait.

Q3. Compute appropriate significance levels for your association tests in Q5. Try out several methods and compare / discuss your results.

Q4. Perform a gene-gene interaction analysis for the quantitative trait, using the tools provided by the "SNPassoc" package. Do you observe any interesting signals?

Q5 Build a random forest. Which variables are important? Are there clusters of variables that are important and do you see any correspondence with the results obtained in Q4? Is this to be expected?

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

Due date: 14 December 2010