

HOWTO: Loading Genotype Data

Gregory Warnes
gregory_warnesurmc.rochester.edu,
Nitin Jain
nitin.jainpfizer.com

October 28, 2009

1 Introduction

This document demonstrates how to use the *GeneticsBase* package to generate marker summary tables *for studies with a small number of markers*. It is written as a step-by-step tutorial. For additional details on each of the R functions utilized, please see the individual help pages

Note: The textual displays described here are not suitable for large numbers of markers. They are intended for reviewing detailed information on a small number of markers, such as those in candidate gene studies, or a small set of markers achieving a 'quality' or 'significance' cutoff from a larger set.

2 Example

2.1 Prepare phenotype data

The first step is to prepare the phenotype data. It may be in the form of a SAS dataset, SAS export file, comma-delimited text file (CSV), tab-delimited text file (TSV), or Microsoft Excel spreadsheet file (XLS). It should have one row per observation and one column per variable, and must contain a subject identifier variable that can be used to match observations with the corresponding genotype data.

2.2 Prepare genotype data

You also need to store the genetic call data in a file that can be read into R. *GeneticsBase* package accepts genotype data in a variety of formats:

- standard pedigree (ped) format.

a2m	apoe					
50103	5010004	5090005	5090004	2	2	1
2	3	4				
50103	5010005	5090005	5090004	2	2	1
1	3	4				
50105	5010049	5090021	5090022	2	2	1
1	4	4				
50105	5010070	5090020	5090019	1	2	1
1	3	4				

- **hapmap format** : The hapmap .ped format is a variant of the standard pedigree format. A portion of the first two lines of the hapmap file for chromosome 1 are shown below:

- Pfizer format: First few lines of an example file in Pfizer's data format are shown below:

- Perlegen format: A portion of first two lines of data in the Perlegen format are shown below:

2.3 Load the genotype and phenotype data

Various types of data can be loaded by `readGenes` function. Example files are provided in `data` subdirectory of the `GeneticsBase` package. To access these execute

Supported file formats include:

The Alzheimer's example dataset is stored in the Fbat variant of the .ped Pedigree Format. As it does not include phenotype data, we only use the `genotypc` filename and file type arguments:

The CAMP example dataset is from the ‘Childhood Asthma Management Program (CAMP)’ and includes both genotype and phenotype information. It can be loaded by:

- HAPMAP .ped format

```
> hapmapchr1 <- readGenes(gfile = "hapmapchr1.ped", gformat = "hapmap")
```

```
> PfizerExample <- readGenes.pfizer("PfizerExample.txt", format = "Listing")
```

- Perlegen data format

```
> PerlegenExample <- readGenes("PerlegenExample.txt", gformat = "perlegen")
```