HOWTO: Loading Genotype Data

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

April 30, 2008

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 packge 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 chrmosome 1 are shown below:

```
rs2298011 rs1320571 rs11721 rs4018608 rs6685064 rs604618 ....
1347 14 0 0 1 1 1 1 3 3 3 3 1 1 2 2 3 3 3 3 2 2 3 3 2  .....
```

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

```
Locus,Gene Marker,Locus Start,Project,Protocol,Sample ID,Donor ID,Genotype
A1A,C1556G,-1243,P234,103,1028022.1,1028022,G/G
A1B,T127A,20141,P234,103,1028022.1,1028022,A/T
A1B,T5094A,102358,P234,103,1028022.1,1028022,A/T
A1A,C1556G,-1243,P234,103,1035130.1,1035130,G/G
```

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

```
snp_id  genotype
753527  rrrrrrrrrrrrrrrrhrrrrrrrrrrrrrrrrrrrrrrrrr......
752848  hhahaararhhhahrhhhahaharhaahhahrhhahhahnhha......
```

2.3 Load the genotype and phenotype data

In GeneticsBase, both genotype and phentype data is loaded by a single function *readgenes*. This function *readGenes* has four primary arguments: *gfile*, *gfromat*, *file*, and *pformat*. Arguments *gfile* and *pfile* are the names of files containing the genotype and phenotype data (respectively), and arguments *gformat* and *pformat* are the corresponding file formats for 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

```
> library(GeneticsBase)
> setwd(file.path(.path.package("GeneticsBase"), "data"))
```

Supported file formats include:

- fbat .ped format

  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 genotyep filename and file type arguments:

  ```
  > ALZH <- readGenes(gfile = "ALZH.ped", gformat = "fbat.ped")
  ```

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

  ```
  > CAMP <- readGenes(gfile = "CAMP.ped", gformat = "fbat.ped", pfile = "CAMPZ.phe",
  +     pformat = "fbat.phe")
  ```

- HAPMAP .ped format

  A subset of the data for the International HapMap project is available in the hapmap example data set. This file can be loaded via:

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

- Pfizer genetics data format

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

- Perlegen data format

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