Differences between snpStats and snpMatrix

David Clayton

April 25, 2023

The `snpMatrix` and `snpStats` packages

The package "`snpMatrix`" was written to provide data classes and methods to facilitate the analysis of whole genome association studies in R. In the data classes it implements, each genotype call is stored as a single byte and, at this density, data for single chromosomes derived from large studies and new high-throughput gene chip platforms can be handled in memory by modern PCs and workstations. The object–oriented programming model introduced with version 4 of the S-plus package, usually termed "S4 methods" was used to implement these classes.

`snpStats` initially arose out of the need to store, and analyse, SNP genotype data in which subjects cannot be assigned to the three possible genotypes with certainty. This necessitated a change in the way in which data are stored internally, although `snpStats` can still handle conventionally called genotype data stored in the original `snpMatrix` storage mode. `snpStats` currently lacks some facilities which were present in `snpMatrix` (although, hopefully, the important gaps will soon be filled) but it also includes several new facilities. This vignette simply describes differences for users converting from the old `snpMatrix` package.

Classes

Function names have, for the most part, remained unchanged so that existing analysis scripts will continue to work with minimal modification. Initially it was hoped also to maintain the old class names since the classes were (mostly) backwards-compatible. But this proved troublesome and, in versions 1.1.4 and later, the class names have been changed (see Table). Two functions have been provided to help users convert objects of a `snpMatrix` class to the corresponding `snpStats` class:

- `convert.snpMatrix`: Converts a `snpMatrix` object to the corresponding `snpStats` class
- `convert.snpMatrix.dir`: Converts all saved `snpMatrix` objects in a given directory

snpMatrix class	snpStats class
snp.matrix	SnpMatrix
X.snp.matrix	XSnpMatrix
single.snp.tests	SingleSnpTests
single.snp.tests.score	SingleSnpTestsScore
snp.tests.glm	GlmTests
snp.tests.glm.score	GlmTestsScore
snp.estimates.glm	GlmEstimates
imputation.rules	ImputationRules

Table 1: Changes in class names

Differences

A major difference is that the basic class, now SnpMatrix, supports uncertain genotypes, as generated by imputation programs. Two classes have been removed, namely the snp and X.snp classes. These were originally devised to support a loss of dimension of a snp.matrix or X.snp.matrix due to selection of a single row or column with drop=TRUE in force in the selection operator []. However these classes were never fully satisfactory and were seldom used. In snpStats the drop= option is no longer allowed during row and column selection; dimensions are never dropped. A word or warning, however: in the event that drop= does occur in the selection operator, this will force the object to be regarded as a simple matrix of type raw; this is the class that SnpMatrix extends and this class does allow drop=.

There has been a cosmetic, but important, change in the XSnpMatrix class as compared with its forerunner. The Female slot has been renamed as diploid to emphasize that this class is not only used for SNPs on the X chromosome, but for any SNP genotypes which may be haploid; this includes SNPs on the Y chromosome and mitocondrial SNPs.

The functions for computing pairwise linkage disequilibrium statistics have been replaced by a rewritten single function, ld. The large band matrix which this function generates in one usage is stored using the dsCMatrix class defined in the Matrix package, (which is now required).

The function read.pedfile has been rewritten, this time entirely in R. It has different arguments from the function of the same name in snpMatrix and may be somewhat slower, but is somewhat more flexible.

The ImputationRules class has changed as a result of the introduction of the new storage convention for uncertain genotypes. In the new coding, uncertainty of calls is represented by (grouped) posterior probabilities of assignment to the three genotypes. This change was necessary because one of the imputation methods of in snpMatrix only produced a posterior expectation of the genotype (when coded 0, 1 or 2) and this could not be accomodated unambiguously in the extended coding.

The GlmTests and GlmTestsScore classes (formerly snp.tests.glm and snp.tests.glm.score) have changed slightly in order to accomodate ongoing work on methods for multinomial and multivariate phenotypes. The test.names slot has been renamed as snp.names and its

function has been changed slightly (although this should only affect more complicated uses of `snp.rhs.tests`). A new slot, `var.names` has been added; this holds the name of the variable(s) tested against SNPs.