Developments in the snpMatrix package

David Clayton

Diabetes and Inflammation Laboratory
Cambridge Institute for Medical Research
Cambridge University, U.K.

16/11/2010
The original aims

- Developed for one of the first GWA studies (WTCCC)
- Implements an efficient storage mechanism for genome-wide SNP data (1 byte per SNP genotype)
- A set of useful statistical tools:
  - LD statistics
  - Single SNP association tests (+ stratification)
  - GLM-based score tests
Recent developments \textit{(snpMatrix2)}

- Extension of storage mechanism to represent \textit{uncertain} genotype assignments
  - Mainly used for \textit{imputed genotypes}, for example using programs such as MACH, IMPUTE (or for storage of internally imputed genotypes)
  - These are represented as posterior probabilities: $p_{AA}, p_{AB}, p_{BB}$
  - Space divided into 253 regions and still stored as 1-byte RAW variables
- Adaptation of existing statistical methods
- New statistical methods
An uncertain SNP genotype (plotUncertainty)
Adapting statistical methods

- **snpMatrix** relies on *score tests*
- Extension to test methods requires replacement of the indicator variables for *additive* and *dominance* effects by their posterior expectations
- Using existing internal imputation methods, we do this “on the fly” without storing the imputed values
- Small changes to every test routine generalizes this to stored uncertain genotypes
- (There has been a small change to the internal imputation method since one method did not give three posterior probabilities)
New and future statistical methods

In current version:
- Fast GLM estimation routines with SNP genotypes entered either as dependent variable or predictor variables

Under development:
- Hypothesis test functions for multivariate and multinomial phenotypes
- A fast implementation of the (blockwise) LARS algorithm for variable selection

Also:
- Interfaces to read and write data files for the widely-used PLINK toolset