Multiple Imputation of Genotype Data

Below is a brief description of imputing genotype data for pedigree data including the data format. The method here is to perform multiple imputation for one marker or loci at a time.

Data Format

Data should be available in standard linkage format. The data should be provided as a data-frame with the following columns.

Family Id	Individual	Father Id	Mother Id	SNP1	.....	Geno	SNP n	Cov 1	...	Cov n

The individual id must be at least unique within the family. We recommend following the data format described in the kinship package. This will allow the users to generate a sparse kinship coefficient matrix using the method developed in the kinship package. For the genotype imputation it is necessary for the users to specify by marker column of interest using the name 'Geno'. The program will impute genotypes for column that is named as 'Geno'.

Genotype Imputation

In order to impute missing genotypes, we first identify individuals within the pedigree that have genotypes missing. We enumerate every possible combination of genotypes for the given number of individuals. Thus, for three individuals with missing data, there would are 27 combinations of genotypes to consider; for four individuals with missing data, this increases to 81 combinations, etc.

For each enumerated collection of genotypes, a partial likelihood is generated assuming that each of the individuals with a missing genotype has the corresponding genotype in the current collection. The overall partial likelihood for the pedigree is defined as the product of the partial likelihoods calculated for each of the individuals with

missing data. The partial likelihood for a given individual is calculated as follows. We identify the parents of the given individual, or indicate that the individual is a founder. We also identify the children, if any, of the given individual. We calculate the conditional probability of the individual's currently assumed genotype given the genotypes of the parents if the individual is not a founder. Note that this may depend on an assumed genotype of one or more of the parents if they had missing genotypes. If the individual is a founder, we instead use the population frequency of the genotype given the frequencies of the alleles assuming the population is in Hardy-Weinberg equilibrium. For each child of the given individual, we calculate the conditional probability of that child's genotype given his/her parents' genotypes, which includes the current individual of interest. Note that the other parent may have a missing genotype, and the child may also have a missing genotype, so this calculation may depend on one or two assumed genotypes. All of the conditional probabilities calculated from parents and children of the current individual of interest are multiplied together to get a partial likelihood for this individual.

As mentioned above, all of the partial likelihoods for all of the individuals that had missing genotypes are multiplied together to get an overall partial likelihood for the pedigree for the current assumed genotypes of the individuals with missing genotype. This value may, and often will be, zero for a given set of assumed genotypes. We retain only the assumed genotype configurations that had nonzero partial likelihoods. For these genotypes, we then renormalize the partial likelihoods so that they sum to one. In order to impute a set of genotypes for the individuals in the pedigree, we choose a genotype configuration from this collection by drawing from the multinomial distribution defined by this renormalized distribution.

The advantages of performing the imputation in this way are twofold. First, impossible genotype configurations – those that generate Mendelian errors – will not be drawn from the distribution. Second, it is not necessary to take into account the full pedigree information, which could be quite involved for some pedigrees. The disadvantage is that all genotype configurations for all individuals with missing data must be enumerated. The number of such configurations increased exponentially with the number of individuals with missing genotypes, so we could have a very large number of such configurations to consider.

Demo

The function that needs to be used to impute missing genotype is miGenotype. The inputs are the data-frame with the marker for which genotype needs to be imputed named as 'Geno'., number of imputations and the allele-frequency.

Tour of the two functions

```
> demo("mig.tour")
```

Usage of the function for imputing missing genotype

```
>miFilledGeno<-miGenotype(missGenDat,nimp=10,alfreq=0.3)
```

Output

The output object is a data-frame with multiple datasets. If the number of imputations are 10 then the data frame would have 10 times the number of rows as the original dataset. The output data frame has an additional column that indicates the imputation number. This lets the user subset based on the imputation number.