Data structure

Genotype matrix , an matrix, the sample size and the number of features (SNPs).

For individual and locus , , their respective counts of are

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | NA |  |
|  |  |  |  |  |

in which NA is missing data, so for every locus, the effective number of genotype is

**Step 0: Standardization [freq, can be paralleled]**

For individual and locus ,

for outbred population

for inbred

or statistically standardized as

in which is the allele frequency of the reference allele (the fifth column in bim file).

**Step 1: relationship matrix**

There are two grms, denoted as and , respectively

Given individuals, we will generate an matrix , for a pair of individual and

output:

denotes the off-diagonal element of ,

Save as \*.grm.gz, \*.grm.id

**GEAR Command: grmA, can be paralleled, see grmA [package gear.subcommands.grmA]**

it has been realized in “**package** gear.subcommands.wgrmA”.

**Of note**, compared with “**package** gear.subcommands.wgrm”, wgrmA speeds up the code by Collections.binarySearch that helps to find the missing values between two genotype vectors. Similarly, binarySearch has been used for grmA.

Notes: subcommands grm and wgrm are both retired.

**Step 2: principle component analysis (PCA, see pca subcommand in GEAR)**

After PCA of , we have eigenvalue and eigenvector . For , its column is denoted as

**Step 3: linear regression (linear)**

in which to .

We have the data matrix below