GModel and **GModel2**

<No-frills linkage and association mapping software with only a 4-page user manual>

Rex Bernardo
University of Minnesota

Background

GModel and *GModel2* are no-frills computer programs for finding marker-trait associations. The programs can be used for linkage mapping of quantitative trait loci (QTL) in a biparental cross, as well as for association mapping in a germplasm collection. *GModel* requires prior estimates of the proportion of the variation that is due to genetic effects. *GModel2* does not require such prior estimates but requires about 20 times the computation time as *GModel*.

The basic procedure implemented in *GModel* and *GModel2* is as follows:

1. Estimate marker effects for all the SNP loci across the genome, via a genomewide prediction framework.
2. Start with chromosome 1. Adjust the phenotypic values for the genomewide marker effects at the remaining chromosomes. In other words, use all of the genomewide markers on the remaining chromosomes to correct for background effects (which also reflect population structure and kinship).
3. Use stepwise multiple regression to find marker-trait associations on chromosome 1.
4. Repeat steps 2 and 3 for each of the chromosomes.

Details of the method are described in the following articles:

Bernardo, R. 2013. Genomewide markers as cofactors for precision mapping of quantitative trait loci. Theor. Appl. Genet. 126: 999-1009.

Bernardo, R. 2013. Genomewide markers for controlling background variation in association mapping. The Plant Genome doi: 10.3835/plantgenome2012.11.0028.

Files

GModel and *GModel2* run as executable files in Windows (`GModel` and `GModel2`) and in macOS (`GModelmac` and `GModel2mac`). Both *GModel* and *GModel 2* require three input files, plus a parameter file:

1. **Marker names and chromosomes** (Sample: `AMchrom.csv`)
   - This CSV file can be exported from Microsoft Excel or Google Sheets.
   - The marker names should have no spaces or commas, and not be more than 30 characters long.
   - The chromosome names should be numbers, with no letters.
   - The SNPs should be ordered by chromosome (lowest to highest), but the SNPs within a chromosome need not be in map order.

2. **SNP marker data** (Sample: `AMsnp.csv`)
   - The rows correspond to the SNP markers and the columns correspond to the lines or individuals.
   - The order of SNPs in this file should correspond to the order of SNPs in the first file.
   - The SNP genotypes need to be coded as 1 for one homozygote, 0 for the heterozygote, and –1 for the other homozygote.
   - There cannot be any missing data. Imputed marker genotypes with a non-integer code (e.g., 0.5, –0.3) are acceptable.

3. **Phenotypic data** (Samples: `AMphen.csv` for *GModel*, and `AMphen2.csv` for *GModel2*)
   - The rows correspond to the lines or individuals and the columns correspond to the traits.
   - In *GModel2*, the first row gives the name of each trait.
   - In *GModel*,
     - The **first row** gives the estimated proportion (between 0.01 and 0.99) of the trait variation that is due to genetic effects. This estimate can be the entry-mean heritability, or the proportion of the sums of squares that is due to genetic effects from an ANOVA.
     - The **second row** gives the name of each trait.
   - The trait name cannot have any spaces.
   - The order of lines in this file (down the rows) should match the order of lines in the SNP marker data file (across the columns).
   - For each trait, only one record is allowed per line or individual. In other words, if a line was evaluated in, say, 3 locations, the mean of the line across the 3 locations needs to be calculated. The mean is then entered as the single phenotypic record for the line or individual.
   - There cannot be any missing data.

4. **Parameter file** (Samples: `ParmsGModel.csv` and `ParmsGModel2.csv`)
   - This file indicates the number of traits, significance level to be used, names of the 3 input files described above, and name of the output file.
   - The parameter file and the output file need to be in CSV format.
   - Use the sample files as templates. Change only the information in the second column, and rename the file as needed.

Traits	2
SignificanceLevel	0.0000001
SNP&ChromFile	AMchrom.csv
MarkerFile	AMsnp.csv
PhenotypeFile	AMphen.csv
OutputFile	Output.csv

Running the Software in Windows

1. Create a folder for the software, and save the executable files (`GModel.exe` and `GModel2.exe`) and input files in this folder.

2. Double-click on an executable file to run it.

Running the Software in macOS

1. Create a `Desktop/Gmodel` folder, and save the executable files (`GModelmac` and `GModel2mac`) and input files in this folder.
2. Open Terminal (e.g., `Command + Spacebar`, then search for `Terminal`).
3. Type `cd Desktop/GModel`

Steps 3a and 3b apply only the first time you run the software:

3a. In the Terminal, type the following to allow access to the files:

```
chmod u+rwx ~/Desktop/GModel/*
```

3b. Then type each of the following to declare that each program is not a virus:

```
xattr -d com.apple.quarantine GModelmac
xattr -d com.apple.quarantine GModel2mac
```

4. Run the software by typing `./GModelmac` or `./GModel2mac` (don't forget to include `./`)

Practical Matters

- Prior to running *GModel* or *GModel2*, make sure that all of the input files as well as the parameter file are not open in Microsoft Excel or Google Sheets.
- Any adjustments of the phenotypic data need to be done outside of *GModel* or *GModel2*, prior to running either program. These adjustments include calculating means across environments and accounting for missing data.
- Make sure all of the markers are polymorphic.
- A high linkage disequilibrium between two markers makes them redundant, possibly leading to spurious results. If two markers have a high linkage disequilibrium (e.g., $r^2 \geq 0.85$ or 0.90), delete one of the markers. ***SNP-QC*** software is recommended for removing monomorphic and redundant markers.
- Simulation results strongly suggest that for the typical sizes of mapping populations and numbers of SNP markers used in plants, a significance level of 0.0000001 to 0.00001 works best. Don't use a significance level less stringent than 0.0001, particularly if the number of markers is very large.
- The marker effect is the mean effect for the SNP marker allele coded as 1. Suppose the marker effect is –0.15. This indicates that one copy of the allele coded as 1 changes the trait by –0.15 and, conversely, the allele coded as –1 has a per-copy effect of 0.15. So changing from 1 (homozygote) to 0 (heterozygote) increases the trait by 0.15, and changing from 0 to –1 (the other heterozgote) further increases the trait by 0.15.
- The effect and *p*-value for a given marker depend on the other markers that are found significant. This is a natural consequence of multiple regression with non-independent markers. Suppose a significance level of 0.00001 leads to SNP1 and SNP2 as having significant effects, and SNP1 has an effect of 0.15 and a *p*-value of 0.0000005. Further suppose that a subsequent analysis is conducted at a more stringent significance

level of 0.000001, and only SNP1 has a significant effect. In this situation, the estimated effect of SNP1 might no longer be 0.15 and the *p*-value might no longer be 0.0000005.

- The *p*-values in the output file are precise up to seven decimal points. Any *p*-values smaller than 1×10^{-7} are output as zero.
- An analysis may take a few seconds to a few hours, depending on the sizes of the data sets.