

Simulation of 2-way and 3-way epistatic interactions with software GAMETES

1. Single Nucleotide Polymorphism

The human genome is organized in 23 pairs of chromosomes. Genetic markers are *loci* with characterized locations along the genome. They allow to observe the variation of the DNA within a population of individuals. One of the most popular classes of genetic markers is the class of Single Nucleotide Polymorphisms (SNPs). SNPs are biallelic, which means that only two variants (among the four nucleotides A, C, T and G) are observed at a SNP's location on the genome. For example, if B and b are the two variants observable for a given SNP, BB, bb, Bb and bB will be the four combinations that may be observed when one considers the two *loci* in the pair of *phased* chromosomes. However, genotyping technologies do not allow to distinguish between Bb and bB. To measure the DNA variation between two individuals, it therefore suffices to code BB, bb and Bb/bB with three integers. If variant B has the highest frequency in the observed population for this SNP, a widely used coding scheme assigns 0 (major homozygous), 1 (heterozygous), and 2 (minor homozygous) to BB, bb and Bb/bB, respectively.

The genetic markers simulated by the software program GAMETES are coded according to the abovementioned code.

2. Input parameters

To simulate 2-way and 3-way interactions, we relied on the Java package GAMETES (Urbanowicz,R.J. *et al.*, 2012), version 2.1. For 2-way and 3-way interactions, GAMETES requires the following input parameters:

- penetrance table (to be further defined),
- total number of simulated SNPs
- total number of SNPs involved in the epistatic interaction (2 for a 2-way interaction, 3 for a 3-way interaction), hereafter called influential SNPs,
- minor allele frequency (MAF) for each influential SNP (frequency of variant b in the example of section 1),
- interval of MAFs for non-influential SNPs,
- number of cases and number of controls,

- number of replicates, that is number of datasets generated under the above conditions.

In this usage, GAMETES automatically computes the heritability h^2 and the disease prevalence $P(D)$. The heritability is defined as the degree of variation of a phenotypic trait that is explained by the genetic variation between individuals of a given population. The prevalence is measured as the total number of cases in a given population.

The penetrance represents the percentage of affected individuals that share the same given genotype.

Since a SNP presents 3 possible variants, the penetrance table has 9 entries for a combination of two SNPs. It has 27 entries for a combination of three SNPs. Each entry of a penetrance table represents a disease risk.

To generate simulated data under the epistasis hypothesis, we relied on models (that is penetrance tables) already used in the literature. Two-way models 1 and 2 were used by Marchini and co-workers (Marchini *et al.*, 1999). Three-way model 3 was used by Zhang and Liu (Zhang, Y. and Liu, J.S., 2007).

The generic forms of models 1, 2 and 3 are shown in Tables 1, 2 and 3, respectively. Therein, α is the baseline effect, while β is the interaction effect. Model 1 is a multiplicative model. Model 2 is a threshold model, where additional disease susceptibility allele (i.e. recessive homozygous) does not further increase the disease risk as in model 1. In model 3, the disease risk is increased when all interacting loci are heterozygous. The disease risk is also increased when the three interacting loci show different genotypes (dominant homozygous, heterozygous, recessive homozygous). However, the risk is less increased than when the three heterozygous variants are present.


Table 1. Model 1

Model 1	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1 + \beta)^2$	$\alpha(1 + \beta)^3$
aa	α	$\alpha(1 + \beta)^3$	$\alpha(1 + \beta)^4$

Table 2. Model 2

Model 2	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1 + \beta)$	$\alpha(1 + \beta)$
aa	α	$\alpha(1 + \beta)$	$\alpha(1 + \beta)$

Table 3. Model 3

Model 3	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	α	α	α	α	α	$\alpha(1+\beta)$	α	$\alpha(1+\beta)$	α
Aa	α	α	$\alpha(1+\beta)$	α		α	$\alpha(1+\beta)$	α	α
aa	α	$\alpha(1+\beta)$	α	$\alpha(1+\beta)$	α	α	α	α	α

3. Instantiations of the penetrance tables

For each experiment, the 2 to 3 influential SNPs were specified to share the same minor allele frequency (MAF). We simulated 4 conditions per model, varying the common influential MAF in $\{0.05, 0.10, 0.20, 0.50\}$.

We simulated 4000 subjects (2000 cases and 2000 controls), and 100 SNPs. Following the indications in (Marchini *et al.*, 1999), and in (Zhang, Y. and Liu, J.S., 2007), we instantiated four penetrance tables per model. These 12 penetrance tables are shown hereafter. The above settings allowed to constrain the heritability and the disease penetrance to the wished values. Across the three models, the prevalence, $P(D)$, is 0.1. Heritability, h^2 , is 0.005 in model 1, and 0.02 in models 2 and 3.

Model 1	P(D)=0.1, h ² =0.005, MAF=0.05		
	BB	Bb	bb
AA	0.098	0.098	0.098
Aa	0.098	0.2989	0.5222
aa	0.098	0.5222	0.9121

Model 1	P(D)=0.1, h ² =0.005, MAF=0.10		
	BB	Bb	bb
AA	0.096	0.096	0.096
Aa	0.096	0.1971	0.2824
aa	0.096	0.2824	0.4047

Model 1	P(D)=0.1, h ² =0.005, MAF=0.20		
	BB	Bb	bb
AA	0.0921	0.0921	0.0921
Aa	0.0921	0.1445	0.181
aa	0.0921	0.181	0.2266

Model 1	P(D)=0.1, h ² =0.005, MAF=0.50		
	BB	Bb	bb
AA	0.0782	0.0782	0.0782
Aa	0.0782	0.1054	0.1223
aa	0.0782	0.1223	0.142

Model 2	P(D)=0.1, h2=0.02, MAF=0.05		
	BB	Bb	bb
AA	0.0958	0.0958	0.0958
Aa	0.0958	0.5331	0.5331
aa	0.0958	0.5331	0.5331

Model 2	P(D)=0.1, h2=0.02, MAF=0.10		
	BB	Bb	bb
AA	0.0918	0.0918	0.0918
Aa	0.0918	0.3192	0.3192
aa	0.0918	0.3192	0.3192

Model 2	P(D)=0.1, h2=0.02, MAF=0.20		
	BB	Bb	bb
AA	0.0836	0.0836	0.0836
Aa	0.0836	0.2099	0.2099
aa	0.0836	0.2099	0.2099

Model 2	P(D)=0.1, h2=0.02, MAF=0.50		
	BB	Bb	bb
AA	0.0519	0.0519	0.0519
Aa	0.0519	0.1374	0.1374
aa	0.0519	0.1374	0.1374

Model 3	P(D)=0.1, h2=0.02, MAF=0.05								
	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0.0942	0.0942	0.0942	0.0942	0.0942	0.6535	0.0942	0.6535	0.0942
Aa	0.0942	0.0942	0.6535	0.0942	0.1469	0.0942	0.6535	0.0942	0.0942
aa	0.0942	0.6535	0.0942	0.6535	0.0942	0.0942	0.0942	0.0942	0.0942

Model 3	P(D)=0.1, h2=0.02, MAF=0.10								
	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0.099	0.099	0.099	0.099	0.099	0.549	0.099	0.549	0.099
Aa	0.099	0.099	0.549	0.099	0.1436	0.099	0.549	0.099	0.099
aa	0.099	0.549	0.099	0.549	0.099	0.099	0.099	0.099	0.099

Model 3	P(D)=0.1, h2=0.02, MAF=0.20								
	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0.0916	0.0916	0.0916	0.0916	0.0916	0.3141	0.0916	0.3141	0.0916
Aa	0.0916	0.0916	0.3141	0.0916	0.1014	0.0916	0.3141	0.0916	0.0916
aa	0.0916	0.3141	0.0916	0.3141	0.0916	0.0916	0.0916	0.0916	0.0916

Model 3	P(D)=0.1, h2=0.02, MAF=0.50								
	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0.0842	0.0842	0.0842	0.0842	0.0842	0.213	0.0842	0.213	0.0842
Aa	0.0842	0.0842	0.213	0.0842	0.095	0.0842	0.213	0.0842	0.0842
aa	0.0842	0.213	0.0842	0.213	0.0842	0.0842	0.0842	0.0842	0.0842

Bibliographical references

Marchini, J. et al. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37,413-417. doi:10.1038/ng1537.

Urbanowicz, R.J. et al. (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*, 5, 16. doi: 10.1186/1756-0381-5-16.

Zhang, Y. and Liu, J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39,1167-1173. doi:10.1038/ng2110.