Simulations

María M. Abad-Grau[1] *, Nuria Medina-Medina[1], Rosana Montes[1] and Fuencisla Matesanz[2]

[1] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada, Granada 18071, Spain
[2] Instituto de Parasitología y Biomedicina López Neyra
Consejo Superior de Investigaciones Científicas
Granada, Spain

1 Simulation studies

Simulation setup. We have performed simulation analyses using genotype datasets of family trios with the intention to compare performance between $\chi^2 - T_{mhet}$ and Monte Carlo T_{mhet}. We have tried to reproduce the same simulations used in several works to check accuracy in TDTs[1–3]. Therefore, we have performed simulation analyses using haplotype data sets of 200 nuclear families (family trios with both parents and a child). Association rates were estimated based on 100 replications of the simulations below described [1–3].

Only consecutive or overlapping clusters of SNPs (which are known as sliding windows) will be tested together. In order to have simulations of a cluster as it was suggested by Crawford et al. [4], we assumed that recombination rates within all markers tested is very low, which is equivalent to assume they belong to the same low recombination block[5]. The recombination fraction within blocks (θ_B) for a common population with exponential growing, such as an African population, has been estimated to be 0.000088 [6] and this is the value used in this work.

We also modified the way to introduce a disease mutation compared with other works [1–3]. Instead of considering only one ancestral chromosome with the disease causing mutation, or the improvement of using two ancestral chromosomes [2], a more realistic simulation of inheritance of complex diseases was used, in which the number of disease ancestral chromosomes can change, according to the coalescent model, as any other gene does.

Populations were drawn using msHOT [7], a program for generating samples based on the coalescent model that incorporates recombination. The samples from all the populations were obtained using *trioSampling*, a computer program available at the supplementary website.

In the following subsections, we will describe the details about the simulations and will highlight those departures from the setup commonly used [1–3].

* Corresponding author: mabad@ugr.es

1.1 Locus specificity and sensitivity

Simulations for power (sensitivity), i.e., assuming no recombination between the disease-susceptibility locus and some of the markers, are also similar to those used in several works assuming one-founder disease haplotype [8, 2, 3] with the intention to evaluate the power of different methods, except that SNPs used are assumed to be in high LD, i.e., they belong to the same low recombination block [5].

Regarding the way to generate samples from populations (one for each population), four parameters were taken into account. Table 2 shows the parameters and their values. The first parameter, the relative risk of being homozygous for the risk allele, RR, varies from 2 to 10 in steps of 2 in the simulations. The second parameter is the number of disease loci used: one and two different disease susceptibility loci were considered. The third parameter is the genetic disease model. Affected and non-affected individuals were drawn by considering different genetic models for the one and two disease susceptibility loci [3]: additive, dominant and recessive (only locus) and additive, domAnDom, domOrDom, recOrRec, threshold and modified for two loci, and different relative genotype risks (RR) of having genotype DD, defined as $Pr(disease \mid DD)/Pr(disease \mid dd)$ (one disease locus) and of having joint genotypes DD and EE, defined as $Pr(disease \mid DD, EE)/Pr(disease \mid dd, ee)$ (two disease loci), with d And e being the normal alleles and D and E the disease alleles. Relative risks for all other genotypes are computed based on RR [9, 3] (Table 1).

The four parameter is added in our work in order to check the decay in association rates due to genetic distance. We considered 5 different recombination fractions (θ) from the markers to the disease susceptibility locus, ranging from perfect LD (no recombination) to $\theta = 0.0002$.

Regarding the way to draw populations, it was the same as the one used to test robustness, except that only one parameter has been used [8, 2, 3]: the number of disease-susceptibility loci (one or two). The parameter of recombination fraction introduced in our simulations to choose the markers for the samples, forced us to modify the pattern of population growth in order to simulate LD decay with distance in a more realistic way in the human population [10, 4]. Thus, after the first 50 generations with constant size [8, 2, 3], we increased the number of generations with exponential growth from 100 to 5000 and the present population size to 100000 (predictions with sample size of only 10000 are also consistent with larger and closer sizes to the actual human population [10]). To be more consistent with real populations and complex diseases in which different number of founders can carry the disease loci, we used the coalescent model [11] to draw populations with a variable number of founder haplotypes and population growth as explained above. Any position can be a disease susceptibility locus. Disease founder haplotypes are chosen by selecting one SNP whose mutant allele has frequencies in the interval $[0.2, 0.4]$, in order to mimic a common disease [3]. To generate these population sets and in order to select SNPs from the population at different recombination rates from the disease susceptibility locus, we used msHOT [7].

References

1. Sham, P.C.: Transmission/disequilibrium tests for multiallelic loci. Am J Hum Genet **61** (1997) 774–778
2. Zhang, S., Sha, Q., Chen, H., Dong, J., Jiang, R.: Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet **73** (2003) 566–79
3. Yu, K., Gu, C.C., Xiong, C., An, P., Province, M.: Global Transmission/Disequilibrium tests based on haplotype sharing in multiple candidate genes. Genet Epidemiol **29** (2005) 223–35
4. Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A., Stephens, M.: Evidence for substantial fine-scale variation in recombination rates across the human genome. Nature Genetics **36** (2004) 700–706
5. Daly, M., Rioux, J., Schaffner, S., Hudson, T., Lander, E.: High-resolution haplotype structure in the human genome. Nature Genet **29** (2001) 229–32
6. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., , Cox, D.R.: Whole-genome patterns of common dna variation in three human populations. Science **18** (2005) 1072–79
7. Hellenthal, G., Stephens, M.: mshot: modifying hudson's ms simulator to incorpore crossover and gene conversion hot spots. Bioinformatics **23** (2007) 520–521
8. Lam, J., Roader, K., Devlin, B.: Haplotype fine mapping by evolutionary trees. Am J Hum Genet **66** (2000) 659–73
9. Fan, R.Z., Xiong, M.M.: Linkage transmission disequilibrium test of two unlinked disease loci. Advances and Applications in Statistics **1** (2001) 277–308
10. Kruglyak, L.: Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics **22** (1999) 139–142
11. Nordborg, M. In: Coalescent theory. John Wiley and Songs, Chichester, UK (2001) 179–212

Table 1. Alleles E and D are the high-risk disease alleles at the corresponding disease locus. The relative genotype risk of a given two-locus joint genotype is calculated using the penetrance of the joint genotype of ee and dd as the baseline. For example, the relative genotype risk of having joint genotype eE and dD is defined as $Pr(disease \mid eR, dD)/Pr(disease \mid ee, dd)$. The relative genotype risk of the joint genotype EE and DD is denoted as RR, which varies from 2 to 10 in steps of 2 in our simulations. Source: [3, 9].

	Genotype at disease	Genotype at disease locus 2		
Model	locus 1	dd	dD	DD
Additive	ee	1	$1+\frac{1}{4}(RR-1)$	$1+\frac{1}{2}(RR-1)$
	eE	$1+\frac{1}{4}(RR-1)$	$1+\frac{1}{2}(RR-1)$	$1+\frac{3}{4}(RR-1)$
	EE	$1+\frac{1}{2}(RR-1)$	$1+\frac{1}{2}(RR-1)$	RR
DomOrDom	ee	1	RR	RR
	eE	RR	RR	RR
	EE	RR	RR	RR
DomAndDom	ee	1	1	1
	eE	1	RR	RR
	EE	1	RR	RR
RecOrRec	ee	1	1	RR
	eE	1	1	RR
	EE	RR	RR	RR
Threshold	ee	1	1	1
	eE	1	1	RR
	EE	1	RR	RR
Modified	ee	1	1	RR
	eE	1	1	RR
	EE	1	RR	RR

Table 2. Values used to configure sample parameters used in specificity/sensitivity simulations.

Relative risk	2, 4, 6, 8, 10
Genetic model	additive, recessive, dominant
θ to disease loci	0, 5e-05, 1e-04, 1.5e-04, 2e-04
Haplotype length	1, 2, 4, 6, 8, 10