Simulation of candidate gene data

We simulated 10,000 haplotypes based on a coalescent process.[1] The mutation parameter θ (4 × diploid sample size × neutral mutation rate for the entire locus) equaled 25. In humans, this would imply a region of approximately 25 kb.[2] Mutations were assumed to be randomly and uniformly distributed across the region. We selected five marker SNPs with minor allele frequencies larger than 0.1 in each region. The minimum distance between the SNPs was 5kb. We assumed no recombinations, thereby creating a "haplotype block" of about 20 kb. To ensure that haplotypes tests would be appropriate, the regions had to comprise at least three common (frequency > 2%) haplotypes.

A hundred different regions were simulated. In 50 regions, we randomly selected an additional SNP as the disease mutation. This mutation affected a normally distributed continuous "liability" representing the total effect of all risk factors. The proportion of variance explained by the mutation was on average 1% (effect size was log-normally distributed with SD = 1). The amount of dominance was uniform distributed ranging from –1 (completely recessive) to 1 (complete dominance). The "cases" were the subjects with the 10% most extreme scores in the upper tail of the liability distribution. The "controls" were the subjects with the 10% lowest scores. Computational details can be found elsewhere.[3]

The properties defined by the 10,000 haplotypes were viewed as the "population" characteristics of that region. For each of the 100 regions we drew, on the basis of these characteristics, 1,000 samples of 500 cases and 500 controls. In each sample, we tested whether case-control status was associated with 1) the 5 marker genotypes, 2) the overall haplotype distribution, 3) each of the common haplotypes, and

4) each of the 10 two marker haplotypes. This creates a multiple testing problem with correlated tests. For instance, if there were four common haplotypes the total number of tests equaled 5 + 1 + 4 + 10= 20. This set of 20 tests would be viewed as the “candidate gene” study. These tests within such a “study” are correlated because the same SNP may be tested as a single marker, as part of a multi-marker haplotype or as part of multiple two-marker haplotypes.

References

1. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 1983;23:183-201.

2. Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. Trends Genet. 2002;83-90.

3. Van den Oord EJCG. A comparison between different designs and tests to detect QTLs in association studies. Behav. Genet. 1999;29:245-256.