

HAPLOTYPES IN STUDIES OF GENE \times ENVIRONMENT INTERACTION

P. Kraft^{1,2}, D. Cox¹, R. Paynter¹, D. Hunter^{1,3,4}, I. De Vivo^{1,4}

Departments of ¹Epidemiology, ²Biostatistics and ³Nutrition, Harvard School of Public Health, Boston, USA; ⁴Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, USA.

Population-based case-control studies measuring associations between haplotypes of single nucleotide polymorphisms (SNPs) are increasingly popular, in part because haplotypes of a few "tagging" SNPs may serve as surrogates for variation in relatively large sections of the genome. Due to current technological limitations, haplotypes must be inferred from unphased genotype data. Using individual-specific inferred haplotypes as covariates in standard epidemiologic analyses is an attractive analysis strategy, as it allows adjustment for nongenetic covariates, provides omnibus and haplotype-specific tests of association, as well as haplotype and haplotype \times environment interaction effect estimates. We compare the performance of several analytic strategies using inferred haplotypes in the context of matched case-control data. These strategies include (a) using only the most likely haplotype assignment and (b) the "expectation substitution" approach, which uses the posterior probabilities of each haplotype pair given the observed genotypes. For moderate haplotype relative risks (≤ 2) and relatively uncomplicated haplotype structures, all methods performed comparably well (small bias with appropriately-sized confidence intervals). For larger relative risks and more complicated haplotype structures, the most likely haplotype strategy showed noticeable bias towards the null; the expectation substitution strategy still performed well. An application to progesterone-receptor haplotypes and endometrial cancer illustrates that performance depends on how well the observed haplotypes "tag" the unobserved causal variant. These results suggest that the straightforward "expectation substitution" approach is accurate in the context of most candidate gene studies for complex disease; choice of "tag" SNPs and accuracy of relevant environmental measurement have a much greater impact on effect estimates.