

# **Inference on Haplotype effects in Case-Control Studies Using Un-phased Genotype Data**

Glen Satten, Ph.D.  
Senior Biomedical Research Service  
Centers for Disease Control and Prevention

## **Abstract:**

Developed in conjunction with: Michael P. Epstein PhD, Dept. of Human Genetics, Emory University

A variety of statistical methods exist for detecting haplotype-disease association using genetic data from a case-control study. As such data often consist of unphased genotypes (resulting in haplotype ambiguity), such statistical methods typically apply the Expectation-Maximization (EM) algorithm for inference. However, the majority of these methods fail to perform inference on the effect of particular haplotypes or haplotype features on disease risk. As such inference is valuable, a retrospective likelihood is developed for estimating and testing the effects of specific features of SNP-based haplotypes on disease risk assuming unphased genotype data from a case-control study. The proposed method has a flexible structure that allows, among other choices, modeling of multiplicative, dominant, and recessive effects of specific haplotype features on disease risk. In addition, the method relaxes the requirement of Hardy-Weinberg Equilibrium (HWE) of haplotype frequencies in case subjects, which is typically required of EM-based haplotype methods. Also, the method easily accommodates missing SNP information. Finally, the method allows for asymptotic, permutation-based or bootstrap inference. The new method is applied to case-control SNP genotype data from the Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus (FUSION) Genetics study and identify two haplotypes that appear to be significantly associated with type 2 diabetes.