The University of Chicago
Department of Statistics
Ph.D. Seminar

MAOXIA ZHENG
Department of Statistics
The University of Chicago

"Two Statistical Problems in Gene Mapping"

PLEASE NOTE TIME CHANGE
Wednesday, November 2, 2005 at 4:30 PM
110 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

In the thesis, we present two statistical problems in gene mapping.

First, we consider the problem of linkage disequilibrium (LD) mapping with haplotype block structure. The HapMap Project is providing a great deal of new information on high-resolution haplotype structure in various human populations. We consider LD mapping in case-parent trios under the assumption that haplotype blocks and their common haplotypes have been identified. Multipoint mapping methods for observed variant, which aims to detect the association between typed markers and the trait, and multipoint mapping methods for virtual variant, which aims to detect the untyped variant that are strongly associated with the trait, are proposed. We consider the multipoint information in two stages. In the first stage, we only consider the multipoint information in each haplotype block, and we term the methods " single block association analysis (SBOV and SBVV)". In the second stage, we construct a multilocus likelihood for the entire data set where a background LD model is needed, and we term the methods " multipoint likelihood mapping methods (MOV and MVV)". Simulation studies have been performed to compare single-point, SBOV, SBVV, MOV, and MVV association analysis in terms of power to detect the association and accuracy of localization of causal variant. These comparisons give some general insight into the value of using different amounts of multipoint information for association mapping. The methods are applied to the data set of Daly et al. (2001) as well.

Second, we consider the problem of identification of pairs of variants that explain a linkage result. Suppose after linkage analysis and fine mapping, we have localized a strong signal of linkage to a small region in which many SNPs are genotyped, that are tightly linked and may have substantial LD with each other and with the trait. Our goal is to determine which SNP or a combination of SNPs that can fully explain the observed evidence of linkage to the region. Based on Sun et al. (2002)'s single-SNP approach, we propose a SNP pair approach in which: 1) For any given pair of tightly-linked SNPs in the region, we test the null hypothesis: the pair of SNPs is the sole cause of the observed evidence for linkage to the region; 2) To take into account the uncertainty in estimation of the haplotype frequency at the putative causal loci, we propose a parametric bootstrap procedure to assess the significance; 3) To deal with missing genotypes, the percentages of missing genotypes in different IBD sharing categories are estimated and incorporated into the calculation of test statistics. Both simulation studies and data analysis of Horikawa et al. (2000) data set show that our method can have high power to reject non-causal SNPs, even when they are tightly linked and in strong linkage disequilibrium with the causal SNPs.

My presentation will focus on the first part of the thesis due to time constraint.