*Design and Analysis of Genetic Studies After the Hapmap Project: Session Introduction*

Francisco M. De La Vega, Andrew G. Clark, Andrew Collins, and Kenneth K. Kidd

Pacific Symposium on Biocomputing 11:451-453(2006)

# DESIGN AND ANALYSIS OF GENETIC STUDIES AFTER THE HAPMAP PROJECT

FRANCISCO M. DE LA VEGA
*Applied Biosystems, 850 Lincoln Centre Dr.*
*Foster City, CA 94404, USA*

ANDREW G. CLARK
*Department of Molecular Biology and Genetics, Cornell University*
*Ithaca, NY 14853, USA*

ANDREW COLLINS
*Human Genetics, University of Southampton*
*Duthie Building (808), Tremona Road, Southampton, England*

KENNETH K. KIDD
*Department of Genetics, Yale University School of Medicine*
*333 Cedar Street, New Haven, CT 06520, USA*

A large international effort to define the fine patterns of sequence variation along the human genome will be essentially completed when this article is printed. The aim was to generate a genome-wide validated SNP resource and survey of the patterns of allelic association and common haplotypes useful for designing association studies. The outcome of this "HapMap" project is the genotypes of about four million SNPs in DNA samples of individuals from Africa, China, Utah, and Japan. This project has provided an unprecedented amount of empirical data on the patterns of linkage disequilibrium across the human genome that is fueling a large number of population genetic analyses. Aspects of the effective representation and use of this vast data resource in the design and implementation of cost-effective and more powerful disease association studies are among the topics of the eight papers included in this volume.

Three papers deal with the issue of how many tagSNPs will be needed to perform a genome-wide association scan without excessive loss of power compared to doing the scan will all the HapMap SNPs. The paper by deBakker et al. examines SNPs in 61 genes involved in DNA repair typed in people from 7 different population groups. The investigators then select "tag" SNPs using TAGGER, and ask a crucial question – how well does this same set of tag SNPs work to perform association testing in each of the 7 population samples. This has been dubbed the "transferability" problem, and until there is empirical

confidence that tag SNPs have this transferability property, application of LD mapping in populations outside the initial HapMap set will be on rather uncertain ground. The good news is that deBakker et al. find that the percentage of SNPs with an $r^2 > 0.8$ to the tag SNPs ranges from 50% to 85% across the population samples.

The question of how many tag SNPs will be necessary to be able to perform genome-wide association testing is tackled by Magi et al. which use the HapMap sample and run it through an $r^2$-binning method as performed by the REAPER algorithm to identify tag SNPs. Then, taking another approach to the problem of transferability, they ask how many additional SNPs would have to be genotyped to attain the same power of an association test, and they find that from 10-35% more SNPs, beyond the minimal tag SNP set from the first population, need to be genotyped in the second population to reach the same power. Gopalakrishnan et al. apply yet a third algorithm, called FESTA, to infer the number of tag SNPs needed for whole genome association testing, and they reach a figure of 294,000 for the European HapMap sample. By picking this subset of SNPs instead of the full HapMap set, they further calculate a lost of power of only 5-10% under a variety of single-gene models for the disease (including dominance, recessiveness and additivity).

Given the vast number of SNPs now available for association studies and the complexity of data analysis it is essential to have software tools that enable both the selection of cost-effective SNP panels that provide high power and tools that assist in determining the optimal approaches to analysis. The SNPbrowser software, described by De La Vega et al., provides a powerful and flexible interface to an embedded database including the HapMap results together with SNP and gene annotations. The software also integrates metric linkage disequilibrium unit (LDU) maps of the genome and step-by-step wizards that implement a number of algorithms for the selection of non-redundant subsets of tag SNPs. The considerable flexibility offered by this tool makes it amenable to the design and implementation of the vast majority of association studies. The paper by Dudek et al. considers in particular how to optimize the analysis of genome-wide association data through the use of their software package genomeSIM. This program is designed to simulate large-scale datasets of population based case-control samples that allow thorough evaluation of alternative analytical methods since the parameters of the simulated disease model are known. While there are several simulation packages available for family-based study designs, the population-based simulation packages are limited to coalescent models which do not accommodate multiple penetrance functions that allow gene-gene interaction to be modeled. As genome-wide association data sets are now beginning to appear,

the importance of modeling gene-gene interactions may soon be revealed and genomeSIM is likely to become an important tool in evaluating the effectiveness of competing approaches to data analysis.

The design and analysis of association studies is still challenging even as high throughput technologies allow the typing of thousands, or even hundreds of thousands of SNPs, since the cost of pursuing false leads after an initial scan needs to be contained. Peter Kraft presents a multi-stage approach where a portion of the samples are genotyped first with a high-throughput genotyping method, and a small number of the most promising variants are then genotyped in the remaining samples with a lower throughput method. The samples sizes in the first and subsequent stages and the corresponding significance levels are chosen to limit the False Positive Report Probability, while maximizing the number of Expected True Positives. Kraft shows that for a fixed budget, the multi-stage strategy has greater power than the single-stage strategy. The expected number of false positives does not change if the true number and effects of causal loci differs from the specified prior, thus limiting the amount of resources spent chasing false leads. On the other hand, Castellana et al. investigate the value of relaxing the rigidity of haplotype block models through a method called “haplotype motifs," which retains the notion of representing haploid sequences as concatenations of conserved haplotypes but abandons the assumption of population-wide block boundaries. They conclude that the benefits of haplotype models are modest, but that haplotype models in general and block-free models in particular are useful in picking up correlations near the boundaries of the detectable level.

Finally, the question of how easy would be to generalize the results of association studies between different populations around the world is addressed by Chen et al. In their paper, a candidate gene association study of the PPAR3 gene on body-mass-index (BMI) of an epidemiological cohort of public school students from Mexico is analyzed. Tag SNPs selected using the HapMap data were used to genotype 1200 subjects. While the present study confirms association between a set of SNPs in LD in the gene and BMI, and the results are promising, the paper discuss the requirements for replication and to discard population stratification effects that may complicate the analysis of associations found in admix populations.

**Acknowledgements**

We would like to acknowledge the generous help of the anonymous reviewers that supported the peer-review process for the manuscripts of this session.