Computational Tools for Complex Trait Gene Mapping: Session Introduction

F. de la Vega, K. Kidd, and A. Collins

Pacific Symposium on Biocomputing 9:89-92(2004)

COMPUTATIONAL TOOLS FOR COMPLEX TRAIT GENE MAPPING

F.M. DE LA VEGA

Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404, USA
E-mail: delavefm@appliedbiosystems.com

K.K. KIDD

Department of Genetics, Yale University School of Medicine
333 Cedar Street, New Haven, CT 06520, USA
E-mail: kidd@biomed.med.yale.edu

A. COLLINS

Human Genetics, University of Southampton
Duthie Building (808), Tremona Road, Southampton, England
E-mail: arc@soton.ac.uk

The mapping of the genes underlying complex traits poses special challenges. The results of several years of effort by many groups in the extension of the linkage mapping methods, used with great effect for localizing major genes, has been disappointing on the whole for complex traits. Now that we have an effectively complete genome sequence and exciting new technologies for genotyping vast numbers of single nucleotide polymorphisms (SNPs) the way is open for the advance of a new strategy. There have already been several successful outcomes for complex trait mapping through the analysis of linkage disequilibrium (LD) and haplotypes. However, these are early days and some of the difficulties are only slowly becoming apparent. Recent evidence [1] suggests that the human genome may contain up to 15 million SNPs. For this reason the probability of actually including a disease causal SNP in a sample of SNPs typed at a spacing of several kilobases is low. Furthermore, this implies that up to 100 other SNPs may be in linkage disequilibrium with a causal SNP. This poses major difficulties for identifying a causal site but the initial target is simply to determine candidate regions with confidence. The International HapMap project [2] has the aim of delimiting haplotype blocks in a number of populations to generate a genome-wide SNP map for association studies. One outcome of this project will be a large body of empirical data on patterns of linkage disequilibrium across the human genome. Other groups and organizations are involved in their own data collection and evaluation studies. Aspects of the effective collection, representation and use of these vast and developing data resources are the topics of the six papers included in this volume.

The potential for whole genome association studies is currently limited by cost. Multiplexing, that is genotyping large numbers of SNPs in parallel per assay, will obviously help reduce costs. The paper of **Sharan et al** shows an algorithmic approach for optimal multiplexing of genotyping assays in generic arrays. Through graph theory this approach partitions SNPs into sets within which every SNP has a unique feature. The results of real data analysis suggest the practical outcomes of such a strategy, permitting, for example, the genotyping of 5,000 SNPs on four all 7-mer arrays.

Whatever system is applied to genotype SNPs concerns over genotyping error, and particularly the effects of error on subsequent analysis, is an ongoing issue. Another concern is the loss of information through 'no call' genotypes – where borderline genotypes are classed as missing. This reduces the error rate but also the number of genotypes returned. In their contribution, **Kang et al** consider the issues of error, no call and missing data and examine the statistical consequences of different scenarios. The basic conclusion is that the benefit of reduced genotyping error rate through not calling certain questionable genotypes is almost exactly balanced by the loss of information due to the reduced number of genotypes. The authors note, however, that in some situations (where one homozygote might be miss-classified as another) no calls might offer greater benefits.

The recognition that a proportion of the genome comprises relatively long blocks of low haplotype diversity [3] was instrumental in the development of the HapMap project. Although there is still controversy about how well the haplotype block model captures the underlying nature of LD in the human genome [4, 5], there have been a number of algorithmic advances in the delineation of blocks since that time. The paper by **Zhu et al** develops a two stage procedure to determine blocks. In this approach a minimum block is extended by the sequential addition of SNPs with the outcome that haplotype blocks are defined in which all SNPs with a minor allele frequency as low as 5% are included. Application to data from four populations reveals that the LD between a SNP and neighboring haplotype blocks is a monotonic function of the distance. This supports the contention that a careful description of the block structure in a region should facilitate mapping. However larger samples are required before instabilities such as decrease in mean block length with increasing SNP density are resolved.

Another area which has understandably seen a recent explosion of interest has been in the determination of haplotypes. Population-based samples pose particular difficulties for reliable haplotype estimation. **Eronen et al** developed a Markov chain approach for reconstruction of haplotypes from multilocus genotypes. This method considers a model that effectively accommodates recombination, motivated by gene mapping in larger regions. Included is a Markov chain model of variable order which uses frequencies of haplotype fragments of different lengths in different regions, thereby accommodating recombination more effectively. The authors used both simulated data and the Daly et al [3] sample to evaluate their methods which outperformed existing methods with sparse maps and were competitive for dense maps. A number of pairwise linkage disequilibrium metrics exist amongst which the absolute value of the D' metric offers a number of advantages. **Kim et al** examine strategies for computing confidence intervals (CI) for D' in order to understand the allele frequency, sample size dependency and the impact on defining haplotype blocks. The authors examined three approaches to developing confidence intervals and concluded that the choice of method was somewhat sample size dependent but there was acceptable coverage (the fraction of times the CI contains the true value of D') for two methods.

Finally, **Bass et al** have developed a software package for generating pedigree data under user-specified conditions. A particular feature is the simulation of variable levels of both recombination and linkage disequilibrium in general pedigrees. The authors recognized a clear need for a program that allows simulation of linkage and association for multiple markers in different data structures from general pedigrees to case and control. It will be important to exploit simulation to examine the statistical properties of the analytical approaches to association/haplotype analysis currently being developed by many groups so the need for such a computational tool is obvious.

The papers in this session illustrate some of the diverse aspects of the exciting, and often controversial, field of complex trait gene mapping. The difficulties involved in performing these types of studies are only now becoming apparent but, fortunately, computational and bioinformatic solutions are keeping pace. It is only a matter of time before the genetic dissection of a number of complex traits is achieved. This will provide the greatly wanted datasets necessary to benchmark novel and more effective computational tools for complex trait gene mapping.

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.

References

1. Botstein, N and Risch N. *Nat. Genet.* 33:228-237 Suppl. (2003).
2. Couzin J. *Science* 296:1391-1393 (2002).
3. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. *Nat. Genet.* 29:229-232 (2001).

4. Wall, J.D. and Pritchard, J.K. *Am. J. Hum. Genet.* 73:502-515 (2003).
5. Stumpf, M.P. and Goldstein, D.B. *Curr. Biol.* 8:1-8 (2003).