

# Computational Issues in Mapping Variation Affecting Susceptibility to Complex Disorders: The Chicken and the Egg

Nancy J. Cox

*University of Chicago, Department of Human Genetics and Department of Medicine,  
507H CLSC, 920 E. 58th Street, Chicago, Illinois 60637*

Received May 2, 2001

Linkage mapping strategies for complex disorders have evolved under a variety of constraints. Some of these constraints reflect the nature of complex disorders and are manifest in limitations on the kinds of data that can be collected, while others were (at least historically) strictly computational. This paper focuses on how computational issues have impacted the design of studies on complex disorders and, conversely, how our study designs have influenced the computational issues that have been addressed. We now have unprecedented computational resources, but also face unprecedented computational and methodological challenges as we move from the linkage mapping of genes influencing susceptibility to complex disorders toward the identification of the actual variation affecting susceptibility to these disorders. The near-term computational and methodological issues we must address will be profoundly influenced by the study designs of the recent past. But future study designs, as well as our investments in computational and methodological research, ought to be developed considering the computational and informatics resources we now have at hand. © 2001 Elsevier Science

**Key Words:** linkage; mapping; algorithms; complex disorders.

## A BRIEF AND SELECT HISTORY OF LINKAGE MAPPING OF GENES FOR COMPLEX DISORDERS

Complex disorders are familial but do not have a simple, Mendelian pattern of transmission. Such disorders are often common, including, for example, diabetes, asthma, cardiovascular disease, and psychiatric disorders, and collectively account for a substantial fraction of public health care expenditures. We remain surprisingly ignorant of the primary defects for many of these disorders, and our efforts to treat and prevent these disorders are hampered by our ignorance. The identification of genetic variation affecting susceptibility to such disorders should increase our understanding of the primary defects. Such knowledge might lead directly to improved

therapies and preventive strategies and might also be useful in designing more sophisticated epidemiological studies that would enable us to identify more specific nongenetic risk factors that might be targets for cost-effective treatment and preventive strategies.

The identification of genes for relatively simple, Mendelian disorders was a straightforward, if sometimes arduous, procedure. We have not been as successful in even the linkage mapping of genes for complex disorders. In some ways, it is premature to judge the success of these endeavors, since the first genome-wide screens in complex disorders including more than 50 families were not published until the mid-1990s. Given the modest linkage signals reported in most of the early studies (on relatively small samples), it is not surprising that fine-mapping and positional cloning studies were not given immediate priority. Instead, much attention has focused

on issues of study design and methods of analysis in an attempt to forge a consensus on an optimal strategy for linkage mapping of genes for complex disorders.

Predictably, this effort to forge a consensus has been an exercise in futility. Optimal strategies for data collection and analysis clearly depend on exactly those aspects of complex disorders of which we remain ignorant. Each disorder may well be unique in its complexity, and a strategy universally recognized as optimal for the linkage mapping of genes for complex disorders remains elusive. There is widespread agreement that the paradigms successfully applied to mapping and then cloning genes for simple, Mendelian disorders must be modified or replaced if we are to be similarly successful in complex disorders. But there is much less agreement on what should be done. Key uncertainties include choice of population to be studied, size and structure of pedigrees to be sampled, and methods for genetic analysis. Here, we focus on how computational issues have influenced the choices investigators have made.

## COMMONLY USED ALGORITHMS FOR LINKAGE MAPPING

Although there are a large number of software packages used in linkage studies, there are two main algorithms underlying the calculations. The approach originally described in Elston and Stewart (1971) scales linearly in the number of individuals in the pedigree, but exponentially in the number of markers used in analysis. Thus, it is ideal for linkage studies of large pedigrees, but in its original form was practical for the analysis of only a single marker with disease. A number of improvements to the original algorithm have been suggested (Cottingham *et al.*, 1993; O'Connell and Weeks, 1995), and multipoint analyses using several markers at a time can now be conducted on large pedigrees. The algorithm originally described in Lander and Green (1987) scales linearly in the number of markers but exponentially with the number of individuals in the pedigree. This algorithm has also undergone substantial improvement (Kruglyak *et al.*, 1996; Kruglyak and Lander, 1998; Gudbjartsson *et al.*, 2000; Markianos *et al.*, 2001) and is routinely used for multipoint linkage analysis of tens to hundreds of markers in moderately sized pedigrees (see Markianos *et al.*, 2001, for examples).

The Elston–Stewart algorithm is most commonly used for parametric linkage analysis—i.e., analyses in which the genetic model (defined by the genotype-specific penetrances and disease susceptibility allele frequency) is specified. The Lander–Green algorithm has probably

been more commonly utilized for nonparametric or allelesharing analyses, but is equally well suited for parametric linkage analysis. As noted above, the main practical differences between these algorithms have to do with the size of the pedigree and the number of markers that can be utilized in analysis. The development and extension of both algorithms were shaped in part by the needs of the genetics community, and both have, in turn, had a major impact on the kinds of data and types of analyses that have been conducted.

## HOW COMPUTATIONAL LIMITATIONS INFLUENCE THE DESIGN OF STUDIES ON COMPLEX DISORDERS

If human geneticists must be the ultimate opportunists, the study of complex disorders requires complete pragmatism. We can study only what data we can collect, and for late-onset disorders, we must deal with the challenges of substantial missing information. A primary way to supplement the data that is missing due to key family members being deceased and/or unavailable for study is through the simultaneous consideration of many highly polymorphic markers in multipoint linkage analysis. As noted above, there have been computational limitations in the size of pedigrees for which it is feasible to compute exact multipoint linkage likelihoods. While there are clearly scientific issues that can and do influence the size and structures of families chosen for study, these computational limitations undoubtedly had an affect on study designs as well. At the very least, it is harder to justify collection and genotyping of samples from large pedigrees when that information cannot be fully utilized.

But it should also be noted that the computational challenges on which we have chosen to focus reflect not only the computational limitations inherent in our methods of analysis, but also the prevailing trends and fashions in the mapping methods considered most appropriate for complex disorders. Rapid multipoint calculations for small to moderate-sized families were developed during a period in which there was a marked emphasis on methods applicable to the analysis of affected sib pairs and related approaches most relevant to families in this size range (Risch, 1990a,b; Weeks and Lange, 1988; Kruglyak and Lander, 1998; Kruglyak *et al.*, 1996; Kong and Cox, 1997). The conventional wisdom on pedigree sizes and structures considered most appropriate for genetic analysis of complex disorders does change as rapidly as high fashion, however, and there is ongoing appreciation of the value of large

pedigrees for genetic studies of complex disorders. It is inevitable that the size of the pedigree for which exact multipoint calculations are feasible will increase (e.g., Gudbjartsson *et al.*, 2000; Markianos *et al.*, 2001) and that the quality and speed of approximate calculations of multipoint linkage likelihoods for pedigrees of arbitrary size and complexity will improve as well (e.g., Sobel and Lange, 1996; Heath, 1997).

Our ability to model the complexities that characterize the genetic component to complex disorders has also been compromised by the computational challenges such modeling entails, and this, in turn, has also affected the data collected for our studies. Although we essentially define complex disorders as those arising as a consequence of the actions and interactions of many genetic and nongenetic factors, we have generally tried to map the contributing genetic loci one at a time. Approaches that allow us to include the gene  $\times$  gene and gene  $\times$  environment interactions that almost certainly characterize the genetic component to complex disorders may substantially improve both the power and the resolution of linkage mapping (Cox *et al.*, 1999; Cordell *et al.*, 2000; variance components interaction studies). Unfortunately, data on even the known nongenetic risk factors have been collected only rarely, at least in part because of our inability to analyze such data. Once there are appropriate and computationally feasible approaches for incorporating information on nongenetic factors affecting risk of disease into linkage mapping studies, we can hope that such data will be routinely collected and effectively utilized.

## CURRENT AND FUTURE CHALLENGES IN GENETIC STUDIES OF COMPLEX DISORDERS

The data we are now using to localize genetic variation affecting susceptibility to complex disorders reflects both the nature of the disorders we study and the limitations, current and historical, of our methods and computational resources. The key challenge that we must now address is how to move from merely localizing genes for complex disorders to identifying the genetic variation that actually influences susceptibility to these disorders.

There might well be as many legitimate approaches to solving this challenge as there are unique data sets. Our general approach (Horikawa *et al.*, 2000) places a major emphasis on identifying the genetic variation that can adequately explain the original evidence for linkage. While it is also true that causal variants should show

reproducible association with disease and have measurable biological function and appropriate physiological consequence, these follow-up studies, especially the functional and physiological assays, can be time-consuming and expensive. Thus, we would argue that the data most useful for the initial fine-mapping and positional cloning studies that are the natural follow-up to successful linkage mapping studies are precisely those used in the original linkage mapping studies and that the most efficient strategy for positional cloning of susceptibility genes for complex disorders is to use the original linkage data in conjunction with newly typed SNP data from the same samples to winnow the number of candidate polymorphisms that must undergo the more expensive functional and physiological testing. It is widely appreciated that even when we try to duplicate exactly a study design resulting in successful mapping of a susceptibility locus for a complex disorder, we are unlikely to replicate the result (Suarez *et al.*, 1994). Thus, any data other than those providing the original evidence for linkage have uncertain value in the initial follow-up studies.

Even the initial stages of follow-up studies, including linkage and linkage disequilibrium (LD) fine-mapping of regions of interest, can be quite computationally intensive. Part of the reason for the computational intensity is the sheer amount of information likely to be generated in the context of these fine-mapping studies. For example, in the fine-mapping and positional cloning studies in a region containing the diabetes susceptibility locus *NIDDM1*, Horikawa *et al.* (2000) reported resequencing  $\sim 80$  kbp of contiguous DNA in 10 samples, as well as a comparable amount of DNA scattered over a larger region. More than 200 variant sites were identified (182 variant sites in the contiguous region resequenced) and nearly 100 of these (mostly SNPs) were typed in samples of unrelated patients (from the original genome-wide screen) and randomly ascertained individuals from the same population. The most interesting of these polymorphisms ( $\sim 30$ ) were typed in all members of all families included in the original genome-wide screen. Preliminary studies using these polymorphisms can be rapidly completed—e.g., comparison of allele and haplotype frequencies among groups—but the more powerful approaches, such as multipoint, likelihood-based linkage disequilibrium mapping (McPeck and Strahs, 1999), are quite computationally intensive.

The computational difficulties stem partly from the nature of the information (much of the data consisting of diploid genotypes rather than unambiguously established haplotypes) used in these analyses and partly from the number of markers that are included in analysis.

Application of the latest algorithms for LD mapping using the decay of haplotype sharing (DHS) approach of McPeck and Strahs (1999) requires 1–2 weeks of computational time on a dual processor Sun workstation. And even this is not really a fully optimal analysis. As noted above, in the contiguous 80-kbp region that was completely resequenced, 182 variant sites were identified, but only 96 polymorphisms have been genotyped. While all informative (minor allele frequency > 0.05) polymorphisms having unique patterns in the 10 individuals used for resequencing were genotyped, there were many informative polymorphisms with identical patterns in these 10 individuals. Initial studies on the polymorphisms from this region showed that polymorphisms with informative, identical patterns in the 10 individuals used for resequencing were in near-perfect LD when typed in a larger (> 200 individuals) sample. Eventually, resources were allocated for the typing of only those polymorphisms with unique patterns in these 10 individuals. This strategy is not unreasonable for analyses that will consider a single polymorphism at a time, whether in allele and haplotype frequency comparisons or in studies designed to test the ability of the polymorphism to partition the original evidence for linkage (Horikawa *et al.*, 2000). The LD mapping approaches utilizing the DHS method, however, should ideally be applied to the complete information from a region—the LD among sites that made them appear less informative for simple studies can be critically important in evaluating the haplotypes shared among affected individuals. For the data reported in Horikawa *et al.* (2000), including this additional information would double the number of markers included in the DHS, substantially increasing the computational burden.

Recent studies focused on genotype/phenotype relationships in model organisms suggest that there may be considerable additional molecular complexity relevant to our studies on complex disorders. In these studies, combinations of variants appear to have much larger effects on phenotype than are apparent from the marginal effects of the individual sites (Stam and Lauri, 1996). This general observation is not so different from what is already suspected for loci with known effects on complex disorders, such as those observed for HLA region variation in autoimmune disorders. But such molecular complexity has yet to be adequately incorporated into commonly used genetic models, and approaches that consider the phenotypic effects of multiple sites simultaneously will clearly increase computational burdens substantially, not to mention the challenges that will be generated in assessing significance of observations.

The expense and difficulty of typing SNPs is being reduced at a rapid rate, which should lead to a profound increase in the amount and complexity of data used in fine-mapping and positional cloning studies. Because of the composition of samples collected for linkage mapping in complex disorders, much of this initial data will be generated in unrelated individuals, or affected sib pairs, in which haplotypes cannot be established unambiguously, and thus analyses of these data will be computationally intensive. Moreover, the need for more realistically complex models at all levels of genetic analysis will also tend to increase computational burdens. The methodological and computational challenges we now face may well become the limiting factor in identifying genetic variation affecting susceptibility to complex disorders.

## REFERENCES

- Cordell, H. J., Wedig, G. C., Jacobs, K. B., and Elston, R. C. 2000. Multilocus linkage tests based on affected relative pairs, *Am. J. Hum. Genet.* **66**, 1273–1286.
- Cottingham, R. W., Jr., Idury, R. M., and Shaffer, A. A. 1993. Faster sequential genetic linkage computations, *Am. J. Hum. Genet.* **53**, 252–263.
- Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I., and Kong, A. 1999. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nat. Genet.* **21**, 213–215.
- Elston, R. C., and Stewart, J. 1971. A general model for the genetic analysis of pedigree data, *Hum. Hered.* **21**, 523–542.
- Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., Kong, A. 2000. Allegro, a new computer program for multipoint linkage analysis, *Nat. Genet.* **25**, 12–13.
- Heath, S. C. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models, *Am. J. Hum. Genet.* **61**, 748–760.
- Horikawa, Y., Oda, N., Cox, N. J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T. H., Mashima, H., Schwarz, P. E. H., del bosque-Plata, L., Horikawa, Y., Oda, Y., Yoshiuchi, I., Colilla, S., Polonsky, K. S., Wei, S., Concannon, P., Iwasaki, N., Schulze, J., Baier, L. J., Bogardus, C., Groop, L., Boerwinkle, E., Hanis, C. L., Bell, G. I. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus, *Nat. Genet.* **26**, 163–175.
- Idury, R. M., and Elston, R. C. 1997. A faster and more general hidden Markov model algorithm for multipoint likelihood calculations, *Hum. Hered.* **47**, 197–202.
- Kong, A., and Cox, N. J. 1997. Allele-sharing models: LOD scores and accurate linkage tests, *Am. J. Hum. Genet.* **61**, 1179–1188.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach, *Am. J. Hum. Genet.* **58**, 1347–1363.
- Kruglyak, L., and Lander, E. S. 1998. Faster multipoint linkage analysis using Fourier transforms, *J. Comput. Biol.* **5**, 1–7.
- Lander, E. S., and Green, P. 1987. Construction of multilocus genetic linkage maps in humans, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2363–2367.

- Markianos, K., Daly, M. J., and Kruglyak, L. 2001. Efficient multi-point linkage analysis through reduction of inheritance space, *Am. J. Hum. Genet.* **68**, 963–977.
- McPeck, M. S., and Strahs, A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping, *Am. J. Hum. Genet.* **65**, 858–875.
- O’Connell, J. R., and Weeks, D. E. 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance, *Nat. Genet.* **11**, 402–408.
- Risch, N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models, *Am. J. Hum. Genet.* **46**, 222–228.
- Risch, N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *Am. J. Hum. Genet.* **46**, 229–241.
- Sobel, E., and Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores and marker-sharing statistics, *Am. J. Hum. Genet.* **42**, 315–326.
- Stam, L. F., and Lauri, C. C. 1996. Molecular dissection of a major gene effect on a quantitative trait: The level of alcohol dehydrogenase expression in *Drosophila melanogaster*, *Genetics* **144**, 1559–1564.
- Suarez, B. K., Hampe, C. L., and Van Eerdewegh, P. 1994. Problems of replicating linkage claims in psychiatry, in “Genetic Approaches to Mental Disorders” (E. S. Gershon and C. R. Cloninger, Eds.), pp. 23–46, American Psychiatric Press, London.
- Weeks, D. E., and Lange, K. 1988. The affected-pedigree-member method of linkage analysis, *Am. J. Hum. Genet.* **42**, 315–326.