Research Statement

Karl W. Broman

> *We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.*
>
> —R. A. Fisher

My research concerns statistical problems in genetics and genomics. I have been principally involved in the genetic analysis of complex traits, but have worked on a variety of other biological problems, including the analysis of meiotic recombination, and most recently, the analysis of genome-scale measures of epigenetic marks.

While my research overlaps the fields of statistical genetics, computational biology and bioinformatics, at my core I remain a statistician. The special value of the statistician's point of view is best expressed by the above quote, from Fisher's *Design of Experiments*. Science concerns the inference of general principles on the basis of specific data. While such inductive inference is inherently uncertain, careful account of the role of chance can allow us to quantify the uncertainty in our inferential statements. Methods for data analysis and exploration have been developed in many disciplines; unique to statistics is the use of probability models for the quantification of uncertainty in inductive inference.

My research is driven by collaboration. I want to be useful, and so I generally tackle problems for which I have a direct scientific collaborator with interesting data (or at least the prospect of interesting data). Excellent scientific collaborators are important to me, so that I may ensure that I am answering the relevant questions, and because it is such collaboration that I most enjoy.

My work has three parts: data analysis, methods development, and software development. Data analysis is my passion; I love nothing more than a new set of data, with new questions and new puzzles. Few real problems can be addressed cleanly with old methods, and so new ones must be developed. My efforts in methods development are almost entirely persuaded by the problems and data that I have at hand, and my focus is always on the problem and the data; I have retained little interest in methods for the sake of methods. Finally, I expend a great deal of effort on software development. I enjoy computer programming, but this is again something that I do not for its own sake, but in order to properly analyze data. New data lead to new methods, and new methods require new computer programs.

The primary focus of my research is on the genetic analysis of complex traits. While I have worked on problems in human linkage analysis, including many applications, data

diagnostics and methods for quantitative trait linkage analysis, the bulk of my efforts concern quantitative trait locus (QTL) mapping in experimental crosses. I have been involved in a large variety of applications, largely in the mouse and rat but also in Drosophila and fish, spanning a large range of different traits. My methodological work has focused on the development of model selection approaches for the identification of multiple, interacting QTL: methods for teasing apart the system of interacting loci that underly a complex trait. I have also developed methods to deal with practical issues that have arisen in applications, including the case of a spike in the phenotype distribution (for example, in the consideration of mass of gallstones when some individuals display no gallstones), and analysis of the X chromosome.

Over the past 6 years, I have been developing the R/qtl software, for QTL mapping in experimental crosses. R/qtl is written as an add-on package for the R statistical software, which allows us to take advantage of R's powerful graphics capabilities and extensive library of mathematical and statistical functions. The aim of R/qtl is to make complex QTL mapping methods widely accessible and allow users to focus on modeling rather than computing. It includes a variety of data diagnostics; the ability to perform single-QTL genome scans and two-QTL, two-dimensional genome scans by multiple methods, with possible allowance for the presence of covariates (such as sex, age or treatment); and the fit and exploration of multiple-QTL models. My goals for R/qtl were ambitious, but we are now close to achieving them. The package has grown to be quite large and complex (it contains about 12,000 lines of C code and 17,000 lines of R code). Had I known what I was getting into, I might not have embarked on the project, but I am very glad that I did. A particularly important feature of the software is that it is written so that extensions may be readily implemented, and so the software provides an important platform for the implementation of new QTL mapping methods.

My second major focus has been the analysis of recombination at meiosis. Without recombination, genetic mapping, whether by linkage or association, would not be possible, and so even if the analysis of recombination were not useful, the process would still deserve careful study. I have constructed genetic linkage maps in humans and dogs, have studied individual and sex-specific recombination in humans, and have studied crossover interference in humans and mice. I have been glad to see a rebirth of interest in recombination. Essentially all of the conclusions in my original work on the Marshfield genetic maps and recombinational variation in the CEPH families were confirmed in the analysis of Icelandic families by deCODE Genetics. I am particularly proud of my discovery of a large, common inversion polymorphism on human chromosome 8p, which was revealed by the presence of apparent triple recombination events and confirmed by FISH. The inversion is about 3 Mbp long, and the frequencies of the two orientations in Europeans are about 60:40.

Most recently, in collaboration with Andy Feinberg at Johns Hopkins, I have begun to study epigenomics. Epigenetics concerns information outside of the DNA sequence that is transmitted from a cell to its daughter cells, such as methylation, genomic imprinting, and histone modifications. We are working to scale-up assays for the measurement of

epigenetic marks to the entire genome. I have been focusing particularly on the use of microarrays to measure allele-specific expression: in essence, we put cDNA on a SNP-chip, and look for differences in the signal for the two alleles at a SNP. (Of course, the subject must be heterozygous at the SNP, and the SNP must be exonic.) The key issues have concerned the combination of information from multiple probe sets for a SNP and multiple SNPs in a gene.

Concerning my future research: we are quite close (finally) to understanding the mapping of multiple QTL, in experimental crosses, for a single phenotype. My next aim is to tackle the analysis of multiple related phenotypes, especially regarding the case of microarray gene expression data as phenotypes. The principal goal remains to improve our understanding of the etiology of a complex disease, here expanded to include the system of genes, mRNA levels and other intermediate phenotypes, rather than simply the genetic architecture of a single trait. I am further interested in developing improved methods for the analysis of advanced intercross lines, heterogeneous stock, and multiple-strain recombinant inbred lines (such as the Collaborative Cross), and for association mapping with multiple inbred mouse strains. There is much interest in these approaches, but the analysis methods remain primitive.

In general, the direction of my future research will build upon the foundation of statistical genetics that I have established as my focus area. The particular problems that my work will address will evolve in the context of the collaborative research environment in which I am working. The importance of the biological questions will determine the nature and priorities of my research.

Many of my current collaborations (particularly regarding gene mapping in model organisms) proceed via email. While this has been quite successful for me, local collaborations are more enjoyable. It is easier to become more deeply involved in the day-to-day research, and great ideas often come from the more casual interactions among researchers. And so I will be very glad for the many opportunities that I will have to collaborate with researchers in Madison.