Statistical Genetics

E. A. Thompson, University of Washington.

1 INTRODUCTION

1.1 Overview of statistical genetics

This paper will review the development of statistical and computational methods for model-based inference from genetic data. There are two broad classes of such inferences; the first is the inference of underlying genetic mechanisms, when the relationship structure of the observed individuals or populations is assumed known. The second is the inference of structure, from genetic data whose underlying biological basis is assumed known. Although methods for addressing both types of question date back many years, the wealth of data provided by modern molecular genetic technology is changing both the feasible inferences and the methods for making them.

Within the first class of inferences, one major division is between experimental and natural populations, with the majority of plant and animal studies falling in the former category, and human studies in the latter. In an experimental population, individuals can be chosen for breeding, design is a major issue, and missing data are few. In human genetic studies, problems of missing data and sampling procedures dominate. The other major division is between models to assess the extent of the genetic contribution to a trait only from trait data among relatives, and models which include the effects of specific genes or regions of the genome. In the inference of structure from genetic data, the different models and methods of approach depend primarily on the time scale of the desired inference. Questions may be of relationships among individuals (genealogical inference), of relationships among populations (migration studies and models of coancestry), or of the longer term ancestry of genes, populations or species (evolutionary and phylogenetic inference).

This paper will be divided into three main sections: first, inference of genetic effects from trait data on related individuals; second, inference of structure; third, methods for linkage and association studies involving the effects of actual genes or regions of genome on traits of interest. In this large field, it will be possible only to touch on the key historical developments, the current major issues, and on future directions. References are given primarily to papers which provide an entry point to other literature. In this regard, Rao and Chakraborty (1991) deserves special mention; many of the chapters provide additional review and references to many aspects of statistical genetics.

1.2 Introduction to genetics

To provide a framework, it is necessary to review some basic genetics; Ott (1991) provides more detail. The basis of inheritance is in DNA; each chromosome contains a double helical strand of DNA. A gene is the segment of DNA that codes for a protein or enzyme, affecting a trait observable in individuals. A locus indexes the chromosomal position occupied by this DNA. Attention is restricted to diploid organisms, and genes whose segregation from parents to offspring follows Mendel's first law: every individual has two genes at a locus, and every offspring of the individual independently receives a copy of a randomly chosen one of these two genes. Thus also, for the two genes in an individual, one is a copy of a randomly chosen one of the two in the father of the individual, and the other is a copy of a randomly chosen one of the two in the mother. Loci on the same chromosome may be linked; in the formation of a single offspring, genes at such loci do not segregate independently of each other. Analysis of data at linked loci is an increasingly major part of statistical genetics, but for simplicity we defer considerations of linkage to section 4.

The particular form of a given gene is an allele; some genes may exist in many different forms. At a particular locus B we will designate the possible alleles by $b_1, ..., b_k$, and their population frequencies by $q_1, ..., q_k$, $0 \leq q_i \leq 1$, $\sum_i q_i = 1$. The genotype of an individual is the unordered pair of types of his two genes. For example, at a diallelic locus ($k = 2$), there are three genotypes b_1b_1, b_1b_2, and b_2b_2. In an infinite population, in the absence of distorting forces such as selection or non-random mating, genotype b_ib_i will have population frequency q_i^2, and genotype b_ib_j $(i < j)$ will have frequency $2q_iq_j$. A population with genotypic frequencies of this form is said to be in Hardy-Weinberg equilibrium.

The genotype of an individual affects the phenotype, the observable qualitative or quantitative value of the trait of interest. Genotypes at some loci may affect many traits. Some traits may be affected by the genotypes at many loci, as well as by non-genetic factors. Some genes may act only at a certain stage of development, or only in certain environments. Specification of the relationship between genotype and phenotype is through penetrance probabilities. The name derives from the simplest case, where the genotype, G at a single diallelic locus determines a binary trait value, Y. Individuals are affected ($Y = 1$) or not ($Y = 0$). The genotype-phenotype relationship is then given by the three penetrance probabilities:

$$\alpha(g) \;=\; \Pr(Y = 1 \mid G = g) \;\; \text{for} \quad g = b_1b_1,\; b_1b_2,\; b_2b_2.$$

More generally, penetrance functions express any relationship between genotype and phenotype. They may depend on age, sex, or other covariates. In principle, they could be a function of the genotypes at several interacting loci. They may express risk of disease onset or a probability distribution for a quantitative trait, rather than a probability for a time-invariant qualitative phenotype.

1.3 Pedigrees, models, and likelihoods

Statistical genetics must involve data on related individuals. Such relationships may derive only from membership of a particular subpopulation, population, or even species, as will be considered in section 3, but in much of statistical genetics a pedigree relationship among individuals is specified. A pedigree is determined by specifying the parents of every individual. Data are not normally available on all individuals; relevant ancestral connections are specified by identifying unobserved ancestors. A pedigree is then a listing of the (*individual*, *mother*, *father*) triplets. Every individual should either have both parents specified or neither. The latter are the founders of the pedigree. Relationships between individuals are defined only relative to a pedigree specification; by definition founders are unrelated.

Analysis of trait data on related individuals requires a genetic model. The model must specify the three components introduced above. First, there is the population parameter specification, giving the numbers of loci, numbers of alleles, and allele frequencies. Included also are any associations between alleles or between loci, or other factors affecting the joint genotype probabilities of unrelated founder individuals. Second, the model must specify segregation probabilities. For simple genetic models these are the 50/50 probabilities of Mendel's First Law, but in general the model might include the more general alternative of possible segregation distortion (Elston and Stewart, 1971). For linked loci, the segregation specification will include locus orderings and recombination frequencies (Section 4; Ott, 1991). Third, the model must have a penetrance component—a specification of the relationship between genotypes and phenotypes.

Given a model about which inferences are to be made, we require a likelihood. Using the above framework, the likelihood of a genetic model θ, given data $\mathbf{Y}$ consisting of phenotypes for some subset of the members of a pedigree, may be written as

$$L(\theta) \;=\; \Pr_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{G}} \Pr_\theta(\mathbf{Y} \mid \mathbf{G})\Pr_\theta(\mathbf{G}) \tag{1}$$

where $\mathbf{G}$ is a genotypic configuration on all the individuals of the pedigree. Then

$$\Pr_\theta(\mathbf{G}) \;=\; \prod_{founders\ i} \Pr_\theta(G_i) \prod_{non-founders\ j} \Pr_\theta(G_j \mid G_{m_j}, G_{f_j}) \tag{2}$$

$$\Pr_\theta(\mathbf{Y} \mid \mathbf{G}) \;=\; \prod_{observed\ l} \Pr_\theta(Y_l \mid G_l) \tag{3}$$

where G_i is the genotype of individual i, Y_i is his phenotype, and m_i and f_i are the parents of i. This form shows the three components of a model. Founder genotype probabilities are determined by population parameters, and the final term of (2) by the segregation model, while (3) is determined by the penetrance functions. While the likelihood (1) provides a framework for inference, direct likelihood evaluation is complicated by the summation over all possible genotypic configurations on the pedigree. Ways to achieve this summation for a variety of types of genetic model and for pedigrees of varying complexity have been developed (Elston and Stewart, 1971; Lange and Elston, 1975; Cannings et al. 1978).

The genetic models as specified in this section are sufficient to define the probability of observed data on individuals in a given relationship structure. For population data, the model must specify also the population processes which give rise to the observed data. These include the genetic forces of mutation and selection, and the population components of mating structure and migration. These aspects will be considered further in section 3.

2 INFERENCE OF LATENT GENES

2.1 Polygenes and variance components

While a quantitative trait may be controlled by a single Mendelian locus (Section 2.3) the usual assumption, particularly in the plant and animal breeding literature, has been that the genetic component of a trait results from the additive effects of an infinite number of loci, each of infinitesimal effect. The *breeding-value*, *polygenic value*, or *additive genetic effect*, z_i, in a non-inbred individual, i, is then, in the limit, normally distributed with mean 0, and variance σ_a^2. Since Mendelian segregation occurs at each of the contributing loci, over an entire pedigree, $\mathbf{z} = (z_i)$ is multivariate normal, with mean $\mathbf{0}$ and variance-covariance matrix $\sigma_a^2 \mathbf{A}$ where $\mathbf{A}$ is a known matrix determined by the pedigree structure.

The simplest model for a quantitative trait determined by underlying additive genetic (polygenic) variation is

$$\mathbf{y} \;=\; \boldsymbol{\mu} + \mathbf{z} + \mathbf{e} \qquad (4)$$

where $\mathbf{y}$ is the vector of trait values observed on (some) members of a pedigree, $\boldsymbol{\mu} = \mu\mathbf{1}$ is a vector of means $\mathbf{z}$ is the vector of polygenic values for the observed individuals, and $\mathbf{e}$ is a vector of i.i.d. $N(0, \sigma_e^2)$ residuals. The latent variables $\mathbf{z}$ are not observed; they simply define a variance component, whose variance σ_a^2 is a parameter of interest. Trait correlations between relatives i and j depend both on the relationship (determining A_{ij}, the ij component of $\mathbf{A}$) and on the relative values of σ_e^2 and σ_a^2. This relative value determines also the response to selection: the heritability of a trait is $\sigma_a^2/\mathrm{Var}(Y)$, the denominator being the population phenotypic variance.

In addition to additive genetic effects, there may be interactions between the types of the two alleles carried by an individual at loci contributing to the trait (dominance effects), and there may be interactions between loci (epistasis) (Cockerham, 1954). There may also be effects of (unmeasured) shared environment, and gene-environment interactions (Falconer, 1989). All these effects may be modeled by additional variance components.

In many recent studies, more than a single quantitative trait is of interest. The existence of, or population association between, genes affecting different traits will lead to non-zero covariances between the polygenic values of individuals for different traits. Specifically, if z_{li} is the breeding value for the ith individual for the lth trait Cov $(z_{li},\ z_{mj}) = \sigma_{lm} A_{ij}$ where $\mathbf{S} = (\sigma_{lm})$ is the matrix of genetic variances and covariances of interest. The matrix $\mathbf{S}$ determines the joint evolution of the traits, or their joint response to imposed selection (Falconer, 1989).

The first models for the genetic basis of quantitative traits were developed by Fisher (1918), providing a blending of Mendelian genetics with the biometric approach of analyzing correlations. A modern mathematical treatment of the population genetic aspects of quantitative traits is given by Bulmer (1980), while Falconer (1989) is the best general text directed towards methods for the statistical analysis of trait data.

2.2 Quantitative genetics in experimental populations

The 1950s were a time of great expansion of the use of statistical genetics in animal and plant breeding, using polygenic models for the genetic effects underlying quantitative traits of interest (Kempthorne, 1957). The general model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\mathbf{z} + \mathbf{e} \tag{5}$$

A given observation y_i is additively determined by the fixed effects of observed covariates $\mathbf{X}$ modeled through the vector parameter $\boldsymbol{\beta}$ and by random effects $\mathbf{z}$, the inclusion of a particular effect in observation y_i being specified by the ith row of the incidence matrix $\mathbf{T}$ (Henderson, 1986).

In the 1950s and 1960s, limitations in computing power restricted analyses of (5) to simple designs and few effects. Design was a major issue (Robertson, 1959), and *reduced animal models* were developed to limit the dimension of the vector $\mathbf{z}$. The 1970s and 1980s saw vast changes in computing power available, with a resulting development of approaches to analysis of data on less balanced data sets, in which more of the extended pedigree relationships among individuals are taken into account. *Animal models*, which include a random effect z_i for each individual i, are increasingly used. Henderson (1986) reviews the development of estimation methods, in the context of quantitative genetic analyses in animal breeding.

The model (5) is now often used for a vector of observations relating to several different traits observed on subsets of the members of a very large pedigree. Many observed covariates, and numerous genetic and non-genetic variance components are routinely included. While full likelihood analysis of (5) is possible, estimating the parameters $\boldsymbol{\beta}$ as well as the variance components, the latter are of primary interest and REML (Patterson and Thompson, 1971) is more often used. Although often, as in (4), $\mathbf{T}$ is taken as the identity matrix, the matrices $\mathbf{A}$ and $\mathbf{S}$ providing the variance of $\mathbf{z}$, it is convenient to allow for an arbitrary matrix $\mathbf{T}$, either for reduced animal models or to permit modeling several traits within a single vector equation (Meyer, 1989). There have been numerous developments in computational approaches to analysis of data using the general mixed model. Most have employed some form of EM algorithm, but even within this framework there are alternative implementations (Thompson and Meyer, 1986). Other computational approaches have also been considered (see, for example, Meyer 1989).

In earlier approaches (Fisher, 1918; Cockerham, 1954: Robertson, 1959), the effects of alleles that contribute to a quantitative trait were explicitly modeled. With the emphasis on variance components (Kempthorne 1957) and advances in computational developments for inferences under the general mixed model (Searle et al., 1993), the genes become less evident. With the wealth of genetic markers now available (Section 4.1), the genes are returning to the quantitative genetics of experimental populations (Section 4.5).

2.3 Quantitative traits on human individuals

In human genetics, the analysis of quantitative variation has focussed on the major effects of single loci. For example, the model of Elston and Stewart (1971) is

$$y_i \;=\; \mu \;+\; \nu(G_i) \;+\; e_i \tag{6}$$

where the mean effect, μ, and the i.i.d residuals e_i are as in (4), but the genetic effect depends only on the individual's genotype, G_i, at a single locus. This is still a random effect (genotypes are unobserved), but the parameters are now those of allele frequency and the penetrance function ν. For example, the model might be of a single diallelic locus, the two alleles b_1 and b_2 having population frequencies q and $1-q$. Hardy-Weinberg equilibrium for founder genotype frequencies, and Mendelian segregation within the pedigree is normally assumed. To analyze data under such segregation analysis models, likelihoods are computed, parameters estimated and hypotheses tested as in section 1.3.

An alternative approach has been to use the polygenic model of (4), computing likelihoods for the parameters σ_a^2 and σ_e^2 (Elston and Stewart, 1971) or analyzing the environmental and genetic components of phenotypic correlations between relatives (Rao and Morton, 1980). The mixed model (Lalouel et al., 1983) combines both polygenic and major gene components:

$$\mathbf{y} \;=\; \boldsymbol{\mu} \;+\; \nu(\mathbf{G}) \;+\; \mathbf{z} \;+\; e \tag{7}$$

The mixed model (7) provides a useful way of incorporating residual familial correlation after allowing for the effects of a major locus of primary interest. It also provides a more general alternative against which submodels (4) or (6) can be tested. However, exact likelihood evaluation is often infeasible, for the likelihood is a mixture of as many Gaussian forms as there are feasible genotypic configurations $\mathbf{G}$ on a pedigree (see (1)). Computational approximations have been proposed (Hasstedt, 1982; Thompson and Guo, 1991).

A major disadvantage of these models is that they do not model covariate effects jointly with the genetic effects; the effects of age, sex, and other covariates are normally adjusted for in advance. The *regressive models* approach of Bonney (1986) allows for the effects of genes and of covariates to be estimated jointly. These models often have no exact correspondence with a model that includes an underlying Mendelian genotype, but they provide a flexible and powerful approach to the analysis of data on more complex traits.

While this section has presented models in the context of a quantitative trait, the same models and methods can be applied to qualitative traits, either directly through a penetrance function (3) or through modeling an underlying quantitative liability (Rice at al., 1991). While for a quantitative trait there may be problems of age adjustment, for a qualitative trait with delayed onset, there are issues of censoring. Segregation analysis alone has not done well in resolving complex traits, but data on linked markers (Section 4.1) greatly increases the potential information and is rapidly changing approaches to human genetic analysis (Elston et al., 1989).

While in experimental populations there are issues of design, in human genetic analyses the problems are primarily those of ascertainment. Disease traits are often rare, and families enter a study population because of the number or pattern of affected individuals within them. A valid analysis must condition on those trait related outcomes that caused a given family to enter the study. The general issue is a choice of the level of conditioning, with consequences for the balance between power and robustness. If more specific assumptions are made about the population from which families are selected, and about the sampling process, then more information about the underlying genetic model can be extracted from the observed phenotypes of individuals. If fewer assumptions are made, more of the data must be conditioned upon. An extensive review of ascertainment has been given by Ewens (1991); he provides numerous references to this active research area of statistical genetics.

3 INFERENCE OF STRUCTURE

3.1 Population genetics and genetic variation

To make inferences about population structure from genetic data, a population genetic model is required. The data normally comprise population allele frequencies (Section 1.2), and the process giving rise to variation in allele frequency is random genetic drift. The simplest models for genetic drift consider a population of $2N$ genes. The Wright-Fisher model determines the genes for the next generation by random sampling, with replacement, from the current one. Thus given current allele frequencies $\mathbf{q} = (q_1, ..., q_k)$, the alleles counts in the next generation are multinomially distributed with index $2N$ and parameter $\mathbf{q}$. An alternative is the Moran model, in which, at each step, a randomly chosen gene is duplicated, and independently one is deleted, from the current array of $2N$

genes. With a Wright-Fisher generation consisting of $2N$ Moran steps, the two models provide qualitatively similar results. More details of these models, and of many other aspects of population genetic models, can be found in Ewens (1979).

Under random genetic drift, mean allele frequencies do not change, although populations evolving independently diverge. Other genetic forces such as immigration from an external source and selection (differential survival or fertility of genes of certain allelic types) cause change in mean allele frequency. Migration among populations, or shared selection, decreases divergence among them. Differential selection, in which alleles are differently favored among populations, tends to increase population divergence.

In a finite population, extant genetic variation will eventually be lost. New variation is provided by mutation. The type of genetic variation used in statistical analyses of population structure is that of observable alleles—traditionally, blood type or enzyme variants. The simplest model for mutation is that it is symmetric among potential alleles, but the genetic reality is that certain mutational changes are more probable than others. The first stepwise mutation model was that of Ohta and Kimura (1973), applied to changes in the charge state of proteins. Two models widely used for modeling are the infinite-sites and infinite-alleles models (Ewens, 1979). The latter is a model in which every mutation is to a new, not previously existing allele. The infinite-sites model has more structure, and is used to model DNA sequence data; every mutation which occurs is at a different site (DNA base) and so, if the ancestral sequence is known, a complete record of mutational changes exists in the current sequence ("allele"). In terms of population allele frequencies, clearly both models have the same properties.

Molecular genetic technology is vastly increasing the number and types of allelic variants available for statistical analysis. Alleles defined by length differences of DNA segments have become the usual variants considered (Botstein et al., 1980). Such variation allows definition of loci not within a DNA coding region nor related to a functional gene. Microsatellite markers (Jeffreys et al., 1988) show very high variation, and also very high mutation rates. Data on such loci have become used in statistical genetics, in problems ranging from the identification of individuals to the phylogeny of species. For such loci, classical symmetric mutation models are also inappropriate; mutations to neighboring length classes are the most probable.

The ultimate measure of genetic variation is that based on DNA sequence data. As yet there is little population data on DNA sequence variation, but some mitochondrial sequence data exist. The mitochondrion is a small haploid extra-nuclear DNA organelle about 17,000 base pairs long, which is primarily maternally inherited; somatic cells and mature oocytes contain large numbers of mitochondria, each with the same DNA complement. Mitochondrial data is providing many new perspectives on inferences of evolutionary processes and evolutionary relationships (Avise, 1991). Where it is not contradicted by the data, an infinite-sites model of mutation is appropriate.

The patterns of genetic variation observed among individuals and among populations provide information for inferences about relationships and for estimation of the parameters of the causes of variation. Within a pedigree, Mendelian segregation gives rise to patterns of genetic difference and similarity among relatives. In a finite population, frequencies of variant alleles are subject to random genetic drift; the primary parameter of this process is the population size, N. The forces of selection and migration also modify allele frequencies, and may be the focus of inference. The source of new genetic variation is mutation; the mutation rate, μ, is critical. On a longer time scale, new alleles replace old ones as a species evolves, resulting in fixed differences between populations.

3.2 Relationships among individuals

Given data on the phenotypes of individuals, the statistical basis of relationship estimation is straightforward. A genealogical relationship is defined by a pedigree, and for each locus equations (1) to (3) give the probability of the data on the observed individuals, or the contribution to the likelihood for a given relationship. The overall likelihood of any relationship is the product of the contributions from data on independently inherited traits, and the likelihoods of alternative hypothesized relationships can be compared, provided the genetic model for each trait is known. The major difficulty is not in the computation of likelihoods under a given hypothesis, but in the huge number of alternative genealogical hypotheses among a group of individuals.

Similarities among relatives result from the *coancestry* of their genes, or the fact they they carry genes that are *identical by descent* (IBD) being copies of a single gene in some common ancestor. Genetic data provide information on the underlying patterns of gene coancestry, and the prior probabilities of such patterns parametrize a relationship. Thompson (1986) gives a description of likelihood procedures for estimation of pairwise and joint relationships from genetic data. Two relationships are of particular practical importance. One is the parent-offspring relationship used in paternity testing. In this case, data on the mother are normally available in addition to those on the offspring and the putative father. The other is the identity (or monozygous twin) relationship used in forensic analyses. These two relationships differ from the others in that, under a model disallowing typing error and mutation, exclusion of the relationship is possible. An offspring must share at least one allele with his father. An individual must have the same genotype as a forensic sample which derives from him. The distributional properties of likelihood estimates are thus somewhat different for these two relationships.

While multiallelic microsatellite markers provide much information for relationship estimation, some caution is necessary. The distributional properties of the likelihoods can be sensitive to errors in population allele frequencies. Particularly with regard to forensic and paternity analyses, the population genetic model assumptions can influence results. Estimation of other relationships is generally more robust, but in inferring sibship structure using microsatellite data on juveniles from a natural population, it can be important to allow for the possibility of mutation (Painter, 1995), and in validating sibships in human genetic epidemiological studies given data at numerous such loci it is important to allow for data errors.

With the data now becoming available, there is increasing interest in estimating relationships in natural populations of plants or animal. In some cases, only general patterns of relationship can be estimated; for example, Geyer et al. (1993) estimated the joint pattern of relationship among a group of individuals from multilocus DNA fingerprint data. In other cases, "fertilities" or frequencies of paternity are estimated (Roeder et al., 1989; Smouse and Meagher, 1994). With sufficient data, estimation of the actual relationships becomes feasible, but there is only a limited number of independently inherited (unlinked) genetic loci. While ignoring linkage does not bias the estimation of IBD proportions (Thompson, 1986), correct analysis must allow for the dependence in IBD at linked loci. In principle, the likelihood of any specific genealogical hypothesis can still be considered; practical solutions remain to be investigated. With the increasing information available on the genomes of individuals in many natural populations, and with the challenging statistical and computational problems involved, this is an area in which future development is likely.

3.3 Population structure and genetic distance

In assessing relationships among subpopulations the primary data are the differing frequencies of alleles. For population l, let q_{li} denote the frequencies of alleles b_i at locus B. The probability that alleles selected randomly from populations l and m are of the same allelic type is

$$J_{lm} = \sum_i q_{li} q_{mi} \text{ and the correlation in allelic identity is } C_{lm} = J_{lm}/(J_{ll} J_{mm})^{\frac{1}{2}} \qquad (8)$$

In considering levels of population subdivision, the statistical approach has been to partition the variance in allele frequencies in accordance with an underlying model for coancestry of genes within and between subdivisions. Wright (1951) provided the first approach towards analyzing the genetic structure in this way, while Cockerham (1973) provides a more general framework. Other distance measures have been based on some specific model for change in allele frequencies. Examples are the measures of Edwards (1971) based on a model of random genetic drift, and that of Nei (1972) which derives from a substitution model, appropriate to longer-term divergence between populations or species. The latter measure is $D_{lm} = -\ln C_{lm}$ and is the most widely used measure of population differentiation. Nei (1987) discusses its statistical properties under evolutionary processes. Chakraborty and Rao (1991) provide a review of this and other distance measures used in evolutionary studies.

A model of stepwise mutational change necessitates new measures of divergence between populations, which take into account the length differences of observed alleles, and not only the frequency profiles. Goldstein et al. (1995) and Slatkin (1995) have proposed and analyzed measures of population divergence appropriate to the evolutionary model underlying such variation, and shown that it provides better information for the estimation of evolutionary parameters from microsatellite allele frequencies.

3.4 Relationships among populations

The development of methods for analysis of relationships among populations based on allele frequencies have ranged from estimates of admixture (Glass and Li, 1953), to the analysis of data under explicit migration models (Bodmer and Cavalli-Sforza, 1968) to estimates of within-species phylogenies (Cavalli-Sforza and Edwards, 1967). Although relationships among populations are seldom phylogenetic, tree representations have proven useful both on the small scale (Neel and Ward, 1970) and the large (Cavalli-Sforza et al., 1994).

At the species level, genetic differences are qualitative; there are fixed differences in the DNA sequences. The sampling unit is the species, and phylogenetic relationships are to be inferred. Models and methods of likelihood estimation have been developed for a variety of types of molecular genetic data (Felsenstein, 1981, 1988). Closely connected with estimation of phylogenetic relationships has been inference of variation of evolutionary rates, from early hypotheses of a ”molecular clock” (Zuckerkandl and Pauling, 1962) to recent estimates of rate variation among DNA sites. The text of Li and Grauer (1991) provides an excellent introduction to the models, methods and data available for inference in molecular evolution.

Recent work on phylogenetic inference has focused towards estimates based on a single stretch of DNA sampled in each individual. The distinction between gene trees and species trees then becomes important. The most recent common ancestor of the gene in two individuals of different species may predate speciation. Different segments of the genome, observed in the same current individuals, may have quite different ancestries.

This has become of particular relevance in the study of mitochondrial variation. Inferences of evolutionary history have been based on mitochondrial data both within (Cann et al., 1987) and between species (Hassegawa and Kishino, 1989). The evidence for "mitochondrial Eve" (a common ancestor of all extant human mitochondria; Cann et al., 1987) no more than 200,000 years ago, is quite strong. However, generalization to the remainder of the human genome, and to inferences of human history, are less firmly based.

3.5 Evolutionary parameters of genetic variation

Most models for assessing evolutionary relationships assume selective neutrality, and much of the early focus of analyses of genetic variation was on testing this neutrality (Kimura, 1983). Under neutrality, equilibrium levels of variation in a random mating population size N is determined by $\theta = 4N\mu$, where μ is the mutation rate. Under the infinite alleles model, the probability that two randomly chosen genes are of the same allelic types is $1/(1+\theta)$. Early estimates of θ were based on allele frequencies, as in (8), but Ewens (1972) showed that, under equilibrium assumptions, the number of distinct alleles in a sample is a sufficient statistic for θ. Following Ewens' framework, Watterson (1975) considered the (future) use of DNA sequence data in assessing population variation, introducing the infinite-sites model, under which two random genes are expected to differ at θ sites.

Kingman (1982) introduced a whole new area of population genetic modeling and inference with the *coalescent*, the ancestral tree of the sample of current genes. Hudson (1990) reviews the recent burst of population genetic theory relating to the form of the coalescent under a variety of population models. Likelihood inference under the coalescent framework is computationally intensive, because of the large number of alternative ancestral trees, but methods for the estimation of θ have been developed (Strobeck, 1983; Griffiths and Tavaré, 1994; Kuhner et al., 1995). Estimation of population structure, of recombination at the DNA level, and perhaps of selection will follow (Donnelly and Tavaré, 1995). However, the information in a current cross-section of populations is limited; selection is confounded with demographic history. While the debate as to the proportion of extant variation that is neutral will continue (Fu and Li, 1993), the coalescent approach has provided a much clearer view of what inferences are in principle possible, and what statistics can be used to provide them.

4 INFERENCE USING GENETIC MARKERS

4.1 Genetic markers and linkage

Genetic mapping is area of major current activity in statistical genetics of plants, animals and humans. The genes that affect observable traits are located on chromosomes. As an offspring chromosome is formed, crossover events occur; copying of the DNA switches from one parental chromosome to the other. The simplest assumption is that crossovers occur as a Poisson process (Haldane, 1919). The offspring receives a chromosome, consisting of alternating segments of the parent's two chromosomes. Thus, on an offspring chromosome, the alleles at loci which are close together (tightly linked) derive with high probability from a single parental chromosome, and hence have the same grandparental origin. A recombination is said to have occurred between two loci, if the alleles at those loci on a single offspring chromosome are of different grandparental origin. A recombination is the result of an odd number of crossover events between the loci. For tightly linked loci the recombination frequency, r, is small. For independently segregating loci, r is one half.

Where the data are such that the grandparental origins of alleles in offspring can be directly and unambiguously inferred, the linear ordering of genetic loci can be confirmed (Sturtevant, 1913). Linkage can be detected and recombination frequencies can be estimated (Fisher, 1922). By analyzing the patterns of recombination between a large number of loci in a large number of offspring, a complete genetic map of the chromosomes of the organism can be determined. To achieve this, genetic markers are needed—single loci at which the segregation of alleles is easily observed. To be useful, genetic markers must be polymorphic; different alleles must segregate.

Originally, the majority of markers were blood type or enzyme loci. Now new molecular genetic markers (Section 3.1). can be used to construct a map of the genome of any organism. A dense map of the human genome is close to a reality (Murray et al., 1994) and there are comparable maps for several other plant and animal species. For linkage mapping of marker loci the statistical model is well defined; the difficulties are primarily computational (Lander and Green, 1987). Other methods of marker mapping, for example using radiation hybrids (Boehnke et al., 1991), raise more interesting statistical questions, but here we restrict attention to linkage mapping of trait loci from phenotypic data on individuals.

4.2 Classical human genetic linkage analysis

Without well-designed crosses, grandparental origins of alleles may not be clear. The likelihood is obtained by summing over the possible latent possibilities of grandparental gene origins, or, equivalently, it is a sum over latent genotypes as in (1). The genotype of an individual is now a haplotype pair; a *haplotype* is a specification of the alleles on a single chromosome. The segregation probabilities (2) are functions of the recombination frequencies, r. As for fully informative experimental crosses, the data provide evidence of linkage and estimates of recombination frequencies through the likelihood function, although computation is harder and information is less. Haldane (1934) first developed likelihood methods to detect linkage in human families, and Fisher (1934) considered more fully the likelihood estimation framework for estimating recombination.

The period 1945-1970 saw development of linkage methods, but few changes. Haldane and Smith (1947) formalized the likelihood approach and proposed a Bayesian prior on recombination frequencies leading to a posterior probability of linkage. Smith (1953), using likelihood, considered many alternative family designs for linkage detection. Morton (1955) introduced the LOD score, the likelihood ratio measure which has become the standard criterion for assessing linkage between a trait locus and a marker.

As more markers became available, and the genetic mapping of less simply determined genetic traits was attempted, statistical difficulties with the classical approach have emerged. Markers are themselves linked, and different markers may be typed in different studies of the same disease, making it hard to combine results. Penetrance and other trait model uncertainties complicate interpretation of the LOD score. There have been many studies of the power and robustness of classical human linkage analysis. The text by Ott (1991) provides a full and clear account.

4.3 Linkage disequilibrium and population association

Consider two diallelic loci B_i, $i = 1, 2$ with alleles b_{ij}, $j = 1, 2$, and let the frequency of haplotypes bearing alleles b_{1j} and b_{2l} be q_{jl}. Then the simplest measure of association, or *linkage disequilibrium* is

$$D = q_{11}q_{22} - q_{12}q_{21} \tag{9}$$

If the recombination frequency between the two loci is r, in an infinite random mating population the value of D decreases by a factor $(1 - r)$ each generation. Thus, while linkage does not cause allelic associations between loci, it maintains them. Measures of association can be defined also for multiallelic loci, and for several loci (Weir, 1995).

Sturt (1984) provides a review of methods for analyzing both association and linkage, clarifying the relation between the two. Statistically, there is often low power to detect associations. Genetically, associations have many causes. They arise through population admixture and patterns of individual migration, or other population processes resulting in a stratification of the population. Associations will also exist, in any finite population, as a result of random genetic drift. While, in expectation, any given association between unlinked loci decays rapidly, variances are large, and in small populations a number of associations may be found (Hill and Weir, 1988).

Other genetic causes of association include epistatic interactions between genes. For example, the genetic loci of the HLA system control the human immune system. Other disease loci may be tightly linked to HLA, or HLA variants might be the primary genetic cause of a disease, but more often they will modify the effect of other genetic or environmental causes. Due to the extreme polymorphism of the HLA system, and the many component genetic loci, associations can be readily detected; interpretation is less straightforward. A thorough review of HLA and disease associations was provided by Thomson (1981). Although much more is now known about the HLA system, and about some of the associated diseases, this review provides much information about HLA, and about measures and mechanisms of disease association. Another view of genetic association is provided by the *measured genotype analysis* approach (Boerwinkle et al., 1986) where variation in a multivariate phenotype is partioned into the effects of different candidate loci at which genotypes of individuals are observed.

Although population association is neither a necessary nor sufficient criterion for linkage, associations may reflect historical events whose effects remain due to tight linkage ($r < 0.005$). There are limits to the resolving power of linkage analysis using family data (Boehnke, 1994); a locus cannot be more precisely located than to a region within which no recombination is seen. When a new variant allele arises it will create a certain haplotype, and disequilibrium will be created. With tightly linked loci, decay of the disequilibrium will be slow, particularly while the new allele remains rare. The disequilibria of a disease locus with tightly linked markers provides a reflection of the recombination events occurring in the ancestral coalescent (Section 3.5) of currently observed disease alleles. Kaplan et al. (1995) have recently considered the power of disease gene mapping, using this historical perspective on linkage disequilibrium.

4.4 Marker maps and genome searches

With the new molecular technology since 1980, sufficient marker data have become available for genome-wide searches of genes contributing to quantitative genetic variation to be a practical proposition. In human genetic studies, trait data are often limited. The availability of data on markers throughout the genome can compensate for this, but these data raise new computational and statistical questions. The method of *map-specific multipoint linkage analysis* is often used. That is, the marker map is assumed known, and a log-likelihood difference (or "location score") computed for each hypothesised location of the trait locus relative to the hypothesis that no trait locus is linked to this segment of the the marker map. The interpretation of such location score curves is not straightforward, and the influence of map and trait model uncertainties can be large. For human genetic data, where there are no designed crosses and many individuals may be unobserved, there are also computational problems. Thompson (1995) reviews recent computational developments in multipoint linkage analysis.

For complex traits, where there may be numerous genes affecting a trait, in the same or in different families, and where there are problems of delayed onset, and environmental effects, a method that is robust to trait model assumptions is needed. Genetic heterogeneity causes particular problems, although as more genetic marker data become available, methods for resolving such heterogeneity become practical (Ott, 1991). Computation of the likelihood remains a key tool for estimation of effects and locations, once linkage has been detected, but for linkage detection alternative methods are are becoming widely used. Penrose (1935) proposed the use of data sib pairs for detecting linkage, but formal development of a test is far more recent (Suarez et al., 1978). The test rests on the fact that, regardless of the exact trait model, sibs of like trait phenotype have higher than average probability of carrying IBD genes (Section 3.2) at trait loci. Thus they are likely also to be carrying IBD genes at closely linked marker loci; such genes must be of the same allelic type. An observed excess (relative to the null expectation) of allele sharing between affected sib pairs at a given polymorphic marker locus thus provides evidence of linkage.

The test is not restricted to affected sib pairs. For genes of recessive effect, allele sharing between the two haplotypes of a single affected individual provides evidence of linkage (Smith, 1953). More generally patterns of allele sharing in arbitrary sets of affected relatives provide evidence of regions of the genome at which the individuals share genes IBD and thus provide evidence of linkage (Weeks and Lange, 1988). While the test statistic under absence of linkage has a probability distribution not depending on the trait model, the power of the tests will depend on genetic heterogeneity and many other aspects of the trait. Some of these have been considered by Risch (1990). For linkage detection, for complex traits, these methods and generalizations of them offer a way forward.

4.5 Linkage analysis in experimental populations

In a population in which informative crosses can be designed there is far greater potential for finding the genes contributing to quantitative traits. A major area of current research is the detection and mapping of such QTL (quantitative trait loci). For simplicity we consider a (slightly) idealized case of two inbred lines, having very different values of the quantitative trait of interest. The genetic map with an ordered set of genetic loci B_1, $B_2, ...,$ is known. One line has only allele b_{i1} at locus B_i, and the other only b_{i2}. In a cross between the lines, the offspring all have genotype $b_{i1}b_{i2}$ at locus B_i. If an offspring is crossed back to a member of the first line, then in the next generation at each locus there are two possible genotypes, $b_{i1}b_{i1}$ and $b_{i1}b_{i2}$. The traditional approach for detecting QTL is a regression approach; the quantitative-trait phenotypic means for these two genotypic classes of backcross offspring are compared (Soller et al., 1976).

While this approach permits detection of QTL, it has several drawbacks. The size of QTL effects are confounded with the recombination frequency between trait locus and marker, the approach provides no estimate of position or effect, and testing with multiple markers results in many false positives. A slightly more sophisticated approach is that of Thoday (1961) who considers the chromosome segment between two marker loci. At two adjacent loci, B_i and B_{i+1}, the haplotype received from the $b_{i,1}b_{i+1,1}/b_{i,2}b_{i+1,2}$ parent is $b_{i,1}b_{i+1,1}$, $b_{i,1}b_{i+1,2}$, $b_{i,2}b_{i+1,1}$, or $b_{i,2}b_{i+1,2}$, while the other parent provides always $b_{i,1}b_{i+1,1}$. The effects of QTL can thus be detected by regressing the observed values of the trait on the genotypes at these two flanking marker loci. Such interval mapping provides greater power to detect QTL.

Despite its drawbacks, the regression approach is straightforward, and remained the standard method until the recent burst of activity resulting from the opportunity presented by dense marker maps. Lander and Botstein (1989) combined the human LOD score approach, interval mapping, and latent QTL genotypes, into a method which permits likelihood estimation of the positions and effects of QTL, providing also appropriate LOD-score thresholds that allow for the multiple dependent tests. Conditional upon latent genotype, the trait is assumed to be normally distributed (Section 2.3), but the latent QTL genotypes have probability distributions in each individual conditional on the marker genotypes at the flanking markers. The likelihood computed is that for a single QTL of given effect at a given position. This likelihood is maximized over effect size, to give a profile likelihood as a function of position along the chromosome – a location score in the sense of section 4.4.

This approach has become widely used, and other methods are being built upon it. Where there are several linked QTL, interacting QTL, or fixed effects (Section 2.2), the full likelihood method of Lander and Botstein (1989) can become computationally intensive, particularly where the crosses cannot be designed in such a way that multilocus haplotypes are clear. The methods of Haley et al. (1994) combine the regression approach with explicit consideration of the QTL genotypes; in effect, regression of phenotype is on an "expected genotype" at each QTL position, given the data on flanking marker loci. While this method is not an exact likelihood approach, it combines the advantages of Lander and Botstein's genetic model with the simplicity of the traditional regression approach. Zeng (1994) uses multiple regression methods, and allows also for more complex genetic models for the QTL effects; this work begins to combine the quantitative genetic models of Cockerham (1954) with modern QTL likelihood mapping methods.

5 CURRENT AND FUTURE DIRECTIONS

Two big changes have transformed statistical genetics over the last 50 years. One is the change in computing power, which has affected all of statistics. Fifty years ago, data sets were small, designs necessarily simple, and models often simplistic. Now it is feasible to compute likelihoods on large and complex pedigrees, or for multiple genetic traits and markers, using genetic models which incorporate explicitly the effects of multiple genes. However, this feasibility would be of little use without the availability of multiple genetic markers. Detailed molecular genetic information on individual genomes provides sufficient data to resolve genetic effects.

With the availability of genetic data, a more unified picture of the statistical issues is developing. Inferences of population history and structure from genetic data are now not only possible (Section 3), but can then be used to address genetic questions (Section 4.3). Although the sampling and design issues are different, the statistical models of human genetic epidemiology (Section 2.3) and quantitative genetics (Section 2.2) are united by modern gene mapping methods (Sections 4.4, 4.5). Genome searches for genes affecting complex traits are undertaken both in experimental and human populations.

The computational issues also have a statistical aspect, with the development of Monte Carlo methods for analysis of data under complex highly structured models. Monte Carlo estimation of likelihoods or Bayesian posterior distributions has been used in segregation analysis (Thompson and Guo, 1991), for assessing patterns of relationship among individuals (Geyer et al., 1993), for estimating evolutionary parameters (Griffiths and Tavaré, 1994; Kuhner et al., 1995), for multipoint linkage analysis on human pedigrees (Thompson, 1995), and for genetic analyses in experimental populations (Sorensen et al., 1994). These recent developments promise an exciting future, as the data to address more complex issues, requiring more complex models, become available.

ACKNOWLEDGEMENT

I am grateful for several very helpful discussions with Dr. Joe Felsenstein, who provided many useful references and perspectives. I am grateful also to Ken Morgan and Ruth Shaw, for helpful comments, and to Dr. Simon Tavaré for an advance copy of Donnelly and Tavaré (1995).

Bibliography

Avise, J. (1991). Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. *Annual Review of Genetics* **25,** 45–69.

Bodmer, W. F. and Cavalli-Sforza, L. L. (1968). A migration matrix model for the study of random genetic drift. *Genetics* **59,** 565–592.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: Implications for positional cloning of human disease genes. *American Journal of Human Genetics* **55,** 379–390.

Boehnke, M., Lange, K., and Cox, D. R. (1991). Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics* **49,** 1174–1188.

Boerwinkle, E., Chakraborty, R., and Sing C.F. (1986). The use of measured genotypes information in the analysis of quantitative phenotypes in man. I. Models and methods. *Annals of Human Genetics* **50,** 181–194.

Bonney, G.E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42,** 611–626.

Botstein, D., White, R. L., Skolnick, M. H., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32,** 314–331.

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics.* Oxford: Clarendon Press.

Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* **325,** 31–36.

Cannings, C., Thompson, E. A., and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10,** 26–61.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19,** 233–257.

Cavalli-Sforza, L. L., Menozzi, P., Piazza, A. (1994). *The History & Geography of Human Genes.* Princeton, NJ: Princeton University Press.

Chakraborty, R. and Rao, C. R. (1991). Measurement of genetic variation for evolutionary studies. In *Handbook of Statistics* **8**: *Statistical Methods in Biological and Medical Sciences.* C. R. Rao and R. Chakraborty, (eds), 271–316. Amsterdam: North Holland.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39,** 859–882.

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics* **74,** 679–700.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29,** in press.

Edwards, A. W. F. (1971). Distances between populations on the basis of gene frequencies. *Biometrics* **27,** 873–881.

Elston R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21,** 523–542.

Elston, R. C., MacCluer, J. W., Hodge, S. E., Spence, M. A., and King, R. H. (1989). Genetic Analysis Workshop 6: Linkage analysis based on affected pedigree members. In *Multipoint Mapping and Linkage Based on Affected Pedigree Members.* R. C. Elston, M. A. Spence, S. E. Hodge and J. W. MacCluer (eds.). 93–103. New York: Alan R. Liss.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3,** 87-112.

Ewens, W. J. (1979). *Mathematical Population Genetics* Berlin: Springer-Verlag.

Ewens, W. J. (1991). Ascertainment biases and their resolution in biological surveys. In *Handbook of Statistics* **8**: *Statistical Methods in Biological and Medical Sciences.* C. R. Rao and R. Chakraborty, (eds), 29–61. Amsterdam: North Holland.

Falconer, D. S., (1989). *Introduction to Quantitative Genetics.* (3rd. edition) London: Longman Group Ltd.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17,** 368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22,** 521–565.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52,** 399–433.

Fisher, R. A. (1922). On the systematic location of genes by means of crossover observations. *American Naturalist* **56,** 406–411.

Fisher, R. A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled. *Annals of Eugenics* **6,** 66–70.

Fu, Y.X. and Li, W-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709.

Geyer, C. J., Ryder, O. A., Chemnick, L. G., and Thompson, E. A. (1993). Analysis of relatedness in the California condors from DNA fingerprints. *Molecular Biology and Evolution* **5,** 571–589.

Glass, B. and Li, C. C. (1953). The dynamics of racial intermixture—An analysis based on the American Negro. *American Journal of Human Genetics* **5,** 1–20.

Goldstein, D. B., Linares, A. R., Cavalli-Sforza L. L., and Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139,** 463–471.

Griffiths, R. C. and Tavaré, S. (1994). Ancestral inference in population genetics. *Statistical Science* **9,** 307–319.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8,** 299–309.

Haldane, J. B. S. (1934). Methods for the detection of autosomal linkage in man. *Annals of Eugenics* **6,** 26–65.

Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* **14,** 10–31.

Haley, C. S., Knott, S. A., and Elsen, J-M. (1994). Mapping quantitative trait loci in crosses between outbred lines, using least squares. *Genetics* **136,** 1195-1207.

Hasegawa, M. and Kishino, H. (1989). Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* **43,** 672–677.

Hasstedt, S. J. (1982). A mixed-model likelihood approximation on large pedigrees. *Computers in Biomedical Research* **15,** 295–307.

Henderson, C. R. (1986). Recent developments in variance and covariance estimation. *Journal of Animal Science* **63,** 208–216.

Hill, W.G. and Weir, B. S. (1988). Variances and covariances of squared linkage disequilibrium in finite populations. *Theoretical Population Biology* **54,** 705–714.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* D. Futuyma and J. Antonovics (eds.). **7,** 1-44. Oxford: Oxford University Press.

Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332,** 278–281.

Kaplan, N. L., Hill, W. G., and Weir, B. S. (1995). Likelihood methods for locating disease genes in nonequillibrium populations. *American Journal of Human Genetics* **56,** 18–32.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics.* New York: Wiley.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A,** 27–43.

Kuhner, M. K., Yamatyo, J., and Felsenstein, J. 1995. Applications of Metropolis-Hastings genealogy sampling. In *Progress in Population Genetics and Human Evolution* P. Donnelly and S. Tavaré (eds.). In Press.

Lalouel, J.-M., Rao, D. C., Morton, N. E., and Elston, R. C.. (1983). A unified model for complex segregation analysis. *American Journal of Human Genetics* **35,** 816–826.

Lander, E. S. and Green, P. (1987). Construction of multilocus linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* **84,** 2363–2367.

Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–199.

Lange, K. and Elston, R. C. (1975). Extensions to pedigree analysis: likelihood computations for simple and complex pedigrees. *Human Heredity* **25,** 95–105.

Li, W-H., and Grauer, D. (1991). *Fundamentals of Molecular Evolution.* Sunderland, MA: Sinauer Associates.

Meyer, K. (1989). Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genetique, Selection, Evolution* **21,** 317–340.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7,** 277–318.

Murray, J.C., Buetow, K.H., Weber, J.L., and 24 others. (1994). A comprehensive human linkage map with centimorgan density. *Science* **265,** 2049–2064.

Namkoong, G. and Kang, H. (1990). Quantitative genetics of forest trees. In J. Janich (ed.). *Plant Breeding Review*, 139–188. Oregon: Timber Press.

Neel, J. V., and Ward, R. H. (1970). Village and tribal genetic distances among American Indians, and the possible implications for human evolution. *Proceedings of the National Academy of Sciences (USA)* **65,** 323–330.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106,** 283–292.

Nei, M. (1987). *Molecular Evolutionary Genetics.* New York: Columbia University Press.

Ohta, T. and Kimura, M. (1973). The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genetical Research* **22,** 201–204.

Ott, J. (1991). *Analysis of Human Genetic Linkage.* (2 nd. ed.) Baltimore, MD: The Johns Hopkins University Press.

Painter. I. (1995). Sibship reconstruction without parental information. Paper submitted to 1995 WNAR Student Papers competition.

Patterson, H. D. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58,** 545–554.

Penrose, L. S. (1935). The general purpose sib pair linkage test. *Annals of Eugenics* **6,** 133–138.

Rao, C. R. and Chakraborty, R. (1991). (eds). *Handbook of Statistics* **8** *Statistical Methods in Biological and Medical Sciences.* Amsterdam: North Holland.

Rao, D. C. and Morton, N. E. (1980). Path analysis of quantitative inheritance. In *Current Developments in Anthropological Genetics* J. H. Mielke and M. H. Crawford (eds.). New York: Plenum Press.

Rice, J., Neuman, R., and Moldin, S. O. (1991). Methods for the inheritance of qualitative traits. In *Handbook of Statistics* **8**: *Statistical Methods in Biological and Medical*

Sciences. C. R. Rao and R. Chakraborty, (eds), 1–27. Amsterdam: North Holland.

Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics* **46,** 229–241.

Robertson, A. (1959). Experimental design in the evaluation of genetic parameters *Biometrics* **15,** 219–226.

Roeder, K., Devlin, B., and Lindsay, B. G. (1989). Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics* **45,** 363–380.

Searle, S. R., Casella, G., and McCulloch, C. E. (1993). *Variance Components.* New York: Wiley.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457–462.

Smith, C. A. B. (1953). Detection of linkage in human genetics. *Journal of the Royal Statistical Society, Series B* **15,** 153–192.

Smouse, P. E. and Meagher, T. R. (1994). Genetic analysis of male reproductive contributions in Chamaelirium luteum (L.) Gray (Liliaceae). *Genetics* **136,** 313–322.

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47,** 35–39.

Sorensen, D. A., Wang, C. S., Jensen, J., and Gianola, D. (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetique, Selection, Evolution* **26,** 333–360.

Strobeck, C. (1983). Estimation of the neutral mutation rate in a finite population from DNA sequence data. *Theoretical Population Biology* **24,** 160–172.

Sturt, E. (1984). Analysis of linkage and association for diseases of genetic etiology. *Statistics in Medicine* **3,** 57–72.

Sturtevant, A. H. (1913). The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14,** 43–59.

Suarez, B. K., Rice, J., and Reich, T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. *Annals of Human Genetics* **42,** 87–94.

Thompson, E. A. (1986). *Pedigree Analysis in Human Genetics.* Baltimore, MD: The Johns Hopkins University Press.

Thompson, E. A. and Guo, S-W. (1991). Evaluation of likelihood ratios for complex genetic models. *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **8,** 149–169.

Thompson, E. A. (1995). Likelihood and linkage: from Fisher to the future. *Annals of Statistics:* in press.

Thompson, R. and Meyer, K. (1986). Estimation of variance components: What is missing in EM? *Journal of Statistical Computing and Simulation* **24,** 215–230.

Thomson, G. (1981). A review of theoretical aspects of HLA and disease associations. *Theoretical Population Biology* **20,** 168–208.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7,** 256–276.

Weber, J. L., and Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics* **2,** 1123–1128.

Weeks, D. E., and Lange, K. (1988). The affected-pedigree-member methods of linkage analysis. *American Journal of Human Genetics* **42,** 315–326.

Weir, B. S. (1995). *The Analysis of Genetic Data.* (2 nd. ed.) Sunderland, MA: Sinauer Associates.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15,** 323–354.

Zeng, Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136,** 1457–1480.

Zuckerkandl, E., and Pauling, L. (1962). Molecular disease, evolution, and genic heterogeneity. In *Horizons in Biochemistry,* M. Kash and B. Pullman, (eds.). Pp.189–225. New York: Academic Press.

E. A. Thompson

Department of Statistics

University of Washington

Box 354322

Seattle, WA 98195

U. S. A.