Genetic modelling of complex human disorders

Swapan K. Nath* and Partha P. Majumder

Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India
*Present address: Department of Genetics, Case Western Reserve University, Cleveland, Ohio, USA

Understanding the genetic bases of complex human disorders is one of the major challenges in human genetics today. Because there are various sources of complexity, including genotype–environment interactions, teasing apart the various causes of these disorders may not be straightforward. Genetic modelling of data on families with these disorders provides useful insight into the roles of various putative causal factors. In this paper, we provide two models – epistatic and heterogeneity models – for complex disorders and develop methodology of analysis of family data under these models.

THE tremendous progress in human molecular genetics made within the last decade has considerably eased the task of localization and cloning of genes controlling single-locus traits/disorders in humans. Presently, the limiting factor to localization of genes controlling such disorders is the availability of multicase families. However, for disorders such as hypertension and diabetes which do not exhibit single-locus inheritance patterns, the challenge now is to identify the various causal factors and to localize the underlying gene(s). Since highly polymorphic DNA markers at a reasonable density are currently available on human chromosomes, it is now possible to study cosegregation of these markers with the trait/disorder even without a tentative model of inheritance for the trait/disorder. However, such efforts may not be cost-effective, and in any case 'before undertaking DNA studies ... one would ideally like to infer as much as possible about the genetic basis of a trait on the basis of the pattern of disease incidence in families and populations'[1].

The purposes of this paper are to: (i) describe some cardinal features of a complex disorder, (ii) briefly discuss some important sources of complexity, (iii) describe some plausible genetic models for complex disorders, (iv) provide some statistical properties of these models, and (v) describe statistical methodology for analysing nuclear family data under these models. The emphasis is on theoretical developments; no application of the theory is provided, but references to applications are given.

What a complex disorder is and the need for modelling

Genetics of many human disorders are complex in nature in the sense that they exhibit consistent and significant familial aggregation, and have a genetic component in their aetiologies, but do not exhibit simple Mendelian patterns of inheritance. Often no single pattern of inheritance can explain all observed types of aggregation of such a disorder in families. For example, a complex disorder that shows a high degree of familial aggregation but is not inherited in a simple Mendelian fashion, may result from epistatic interactions of alleles at two or more loci. Even when only two recessive loci epistatically interact in the pathogenesis of a disorder, the vast majority of families ascertained through an affected proband have no other affected member. For example, when the population prevalence of a two-locus recessive disorder is 1/1000, about 78% of nuclear families and about 65% of three-generational extended families of the proband are simplex[2]. These figures increase to about 82% and 78%, respectively, when the prevalence decreases to 1/10000. Segregation analysis of such a multilocus recessive disorder may result in incorrectly inferring that the disorder is incompletely penetrant with a large proportion of sporadics[2]. A recent study[3] has shown that for the purpose of detecting linkage, mis-specification of the two-locus model by a single-locus model does not affect the expected maximum lod-score substantially. However, model mis-specification leads to loss of power of detecting linkage and to biased estimates of the recombination fraction and other segregation parameters[3-7]. Further, no single-locus trait can fit the recurrence risks in relatives when the true model of inheritance is oligogenic[8]. For a two-locus trait, two-trait-locus, two-marker-locus linkage analysis can provide substantially more linkage information than standard one-trait-locus, one-marker-locus analysis[9]. Identification of the correct genetic model of a disorder by segregation analysis is, therefore, not only useful but is also necessary for both genetic counselling and localization of genes.

Sources of complexity

The complexity of a genetic disorder can arise in a variety of ways. For example, a disorder may be determined by the joint action of genes and environment, such as insulin dependent diabetes mellitus (IDDM). This disorder has a variable age at onset. It aggregates in families, but does not segregate in a simple Mendelian fashion from parents to offspring[10,11]. There are also possible environmental effects or effects of other loci

(e.g. HLA) on the expression of this disorder. Although no claim is made that the following list is mutually exclusive or exhaustive, some of the more common causes of complexity of a disorder are:

Variable age at onset. All individuals with the appropriate genotype do not manifest the disorder either at birth or at the same age later in life. The classic example is Huntington's disease for which the age at onset ranges from 6 years to 75+ years with a mode between 41 and 45 years[12].

Reduced penetrance. Some individuals with the appropriate genotype manifest the disorder while some others do not. Such reduced penetrance may be due to random, stochastic factors or due to modification of the susceptible genotype(s) by other gene(s)[13].

Phenotypic heterogeneity. All individuals of same genotype do not manifest the same phenotype. Phenotypic heterogeneity can, however, be artifactual. For example, one of the difficult aspects of studying the genetics of psychiatric disorders relates to phenotype definition. Because of the large number of both major and minor psychiatric diagnoses, a primary problem in conducting genetic studies of psychiatric disorders is knowing which ones to include as affected, and which ones to exclude[14]. Inability to define the phenotype homogeneously may give a false indication of phenotypic heterogeneity.

Allelic/genetic heterogeneity. Different alleles either at the same locus or at the different loci, give rise to the same phenotype. In classical terms, the former type is known as intra-locus heterogeneity and the latter type is known as inter-locus heterogeneity. An example of intra-locus heterogeneity is cystic fibrosis (CF). Of all CF patients, 70% carry the same mutation, the $\Delta F508$ mutation[15], a deletion of a specific codon that causes the CF protein to lack an amino acid phenylalanine at amino acid position number 508. Additionally, more than 300 mutations of the CF gene have been reported. The phenotypic effects of some of these mutations can be distinguished, while of some others seem indistinguishable. For example, the *M348K* mutant allele, which is characterized by a T to A substitution at nucleotide position 1175 in exon 7 of the CF gene leading to a methionine to lysine amino acid substitution, is reported to have a phenotypic effect that is indistinguishable from that of $\Delta F508$ (ref. 16). Neurofibromatosis (NF) is a disease for which there is inter-locus heterogeneity. NF can be of two types[17]. The most common form is the von Recklinghausen type (NF1), which is linked to markers on chromosome 17 (ref. 18). The other rare form is acoustic type (NF2), which is linked to markers on a different chromosome 22 (ref. 19).

Involvement of multiple loci. The disorder is determined by the joint action of genes at more than one locus. The genetic mechanism for a specific form of prelingual deafness may be cited as an example. This disorder manifests itself only in individuals who are recessive homozygotes at two involved loci[20].

Environmental influence. Environment, together with genotype, jointly influence the manifestation of the disorder. Coronary heart disease (CHD) can be cited as an example. The contribution of genetic factors to the development of CHD has been estimated[21,22]. From these studies it appears that there is no single gene locus responsible for CHD. Rather, different environmental and genetic factors jointly act and interact in a highly complex fashion in the pathogenesis of CHD.

Problems in the analysis of family data of a complex disorder

Methods of segregation analysis of family data for determining the mode of inheritance of a disorder are well established when the disorder is primarily determined by alleles at a single locus[23]. Suitable modifications have also been made to take into account incomplete penetrance, variable age at onset, etc. However, when a disorder is not transmitted in a simple Mendelian fashion, that is, when the observed segregation probability/ratio (which is defined as the conditional probability of an affected offspring given a parental mating type) is vastly different from that expected under a one-locus model, the disorder is usually described as 'polygenic' (meaning that the disorder is determined by alleles, each with a small undetectable effect, at a very large number of loci, which act additively to produce the disorder phenotype). Heritability of the disorder is then estimated. However, with the identification of single genes in the so-called complex disorders, the concept of polygenic inheritance is now beginning to be challenged[24]. A recent example is Hirschsprung disease, a complex disorder hitherto considered to be of polygenic origin. Patients with this disease suffer from severe constipation and abdominal distension due to congenital megacolon. A gene for this disease, which was traditionally assumed to be polygenic, has now been successfully localized to the pericentromeric region of the short arm of chromosome 10 (ref. 25). Thus, there is an increasing realization and documentation of the fact that many, so-called complex disorders may really be due to effects of genes at one or a small number of loci. Having rejected the single-locus model of inheritance, no attempt is usually made to analyse the data under multilocus models, assuming the involvement of a small number of loci. This is because of the intrinsic problems associated with multilocus models. For example, under a single locus

model, if the disorder is recessive, then in a family, ascertained through an affected child, if both parents are normal, then both of them are obligate heterozygotes. This fact simplifies family data analysis to a great extent, because in all such families, the segregation ratio is 1/4. Such simplicity vanishes even when two unlinked biallelic loci, with alleles (A,a) and (B,b), are considered. In this case, each normal parent can be of one of the genotypes AaBb, Aabb, aaBb. Thus, a normal × normal mating type may be any one of the genotypic matings AaBb × AaBb; AaBb × Aabb; AaBb × aaBb; Aabb × Aabb; Aabb × aaBb; aaBb × aaBb with segregation ratios 1/16; 1/8; 1/8; 1/4; 1/4; 1/4, respectively. This heterogeneity of mating types and segregation ratios introduce considerable complexity in deriving the likelihood function of a set of phenotypic observations on members of a family, and in carrying out computations. Further, because of the small family sizes of humans, low segregation ratios imply that in a large proportion of families only one affected member (the proband) is usually observed. This, in turn, creates problems in data analysis because the members in these families appear as sporadic (non-genetic) cases and/or the normal individuals mimic incompletely penetrant cases. To avoid the confusion regarding whether the affected individual in a single-case family is genetic (but, chance isolated) or sporadic, it may be more practical to select multicase families. However, even though such a selection procedure enriches for segregation at a few loci, these segregants will be at multiple, independent loci[26]. This implies that there will be an intrinsic heterogeneity among such families which will adversely affect linkage-mapping efforts. Variable age at onset adds further complexity, because in this case an individual may be of the 'affected genotype', but may not have expressed the disorder at the time of study.

Development of multilocus models

A number of human disorders and congenital malformations show strong familial aggregation but do not conform to the expected recurrence risks in sibs, or are not transmitted from parents to offspring in a simple, single locus, Mendelian fashion. Various models have been proposed to describe the way in which gene(s) affect the liability of individuals to a disorder. They range from models of the effects of alleles at a single locus to multifactorial/polygenic models representing the effects of genes at many loci and the effects of environment. The multifactorial model, while descriptive, sheds very little light on possible underlying biological mechanism. Thus, extending simple single locus Mendelian models to more than one locus represents the next logical step for exploring possible genetic mechanisms for diseases which show strong familial aggregation. However, inherent in analysing models based on two loci is the question of the biological and statistical aspects of interactions between alleles at the different loci. Straight additivity is rarely a good assumption for any biological mechanism, while epistatic interaction represents a plausible mechanism for many disorders.

To the best of our knowledge, the first two-locus model proposed for a human disorder – psoriasis, a dermatological disorder – was by Steinberg *et al.*[27]. They performed a simple-minded analysis of data from various mating types and concluded that recessive alleles at two unlinked, autosomal, biallelic loci interacted in the manifestation of the disorder, and that individuals who were recessive homozygotes at both loci were affected. Although psoriasis has a variable age at onset, this fact was not rigorously incorporated in Steinberg *et al.*'s[27] analyses.

Li[28] considered the two-locus recessive homozygosis model and derived many useful theoretical results that included showing that Snyder's ratios can be generalized in a straightforward manner to multiple loci. These results are useful for analysing nuclear family data for a disorder that is expressed at birth, provided that families are ascertained randomly or through an affected parent. He also provided further generalizations and results[29]. However, in practice, for recessive disorders, sampling through an affected offspring, not random sampling, is the generally adopted strategy.

Merry *et al.*[30] performed a theoretical study of a two-locus model for a familial disease. They derived conditions for the existence of a stable equilibrium and showed that it could be used to explain a wide range of disease frequencies and patterns of inheritance. However, even though this study considered some diseases with variable ages at onset, the theoretical investigations were carried out without taking ages at onset into account.

A simple graphical method for testing two-locus models was proposed by Greenberg[31]. However, the testing procedure described by the graphical representation is not a test of significance or fit in the statistical sense. It is rather a test of the consistency of the model with biological parameters, namely, the gene frequencies at the two loci. A maximum likelihood test of the two-locus model for coeliac disease was subsequently developed[32], but are not directly applicable to a disorder with a variable age at onset.

Recently, Neuman and Rice[8] have derived formulae for the recurrence risk to various classes of relatives in terms of penetrances and gene frequencies for two locus models.

Methods for segregation analyses of data on ascertained families in respect of multilocus recessive homozygosis models, have primarily been developed by Majumder *et al.*[20,33]. Further generalizations and applications have also

been made, and properties derived[14-16]. These methods developed are applicable to disorders which are expressed at birth or which have variable ages at onset.

Multilocus models with special reference to the multilocus recessive model

Models considered

Many of the challenging problems in the study of hereditary disorders involve use of mathematical/statistical modelling to describe the transmission of a disorder within families. Many complex disorders have not been amenable to genetic analysis under the assumption of single locus or multifactorial models. The observed familial risks are often inexplicable under any single locus or multifactorial models and also segregation analysis has often not been decisive[8]. Consequently, interest has turned to the consideration of the properties of oligogenic models, i.e. genetic models involving a small number of genes.

It is known that the number of possible models for a multilocus system is large[17], which precludes the exhaustive investigation of all possible models. To understand the behaviour of oligogenic models, we shall first consider the simplest case – two autosomal, biallelic, unlinked loci. Locus 1 has alleles A and a with frequencies p_1 and q_1 $(=1-p_1)$, respectively; the two alleles at locus 2, B and b, have frequencies p_2 and $q_2 (=1-p_2)$, respectively. It is assumed that the underlying population is in Hardy–Weinberg equilibrium with respect to each of the two loci and that there is no linkage disequilibrium between the loci. We also assume that penetrances are equal (say f for epistatic models) for all at-risk genotypes and that only those individuals with an at-risk genotype may become affected (i.e. no phenocopies). In heterogeneity models, the penetrance (say g) of a genotype with respect to both loci is computed from marginal penetrances: $g=f_1+f_2-f_1f_2$, where f_1 and f_2 denote the marginal penetrances at loci 1 and 2, respectively. The genotypes and their corresponding population frequencies for the general two locus model are given in Table 1. The notation used to denote penetrances of the corresponding two-locus genotypes is given in Table 2. Now for example, consider a disorder that expresses itself in an individual if either (s)he is a recessive homozygote at each of the two loci (epistatic model) or is a recessive homozygote at any one of the two loci (heterogeneity model). The penetrances of the genotypes for these models are: epistatic: $h_1=h_2=\ldots=h_8=0$, $h_9=f$; heterogeneity: $h_1=h_2=h_4=h_5=0$, $h_3=h_6=f_2$, $h_7=h_8=f_1$, $h_9=g$. In this study, we have focused on the epistatic and the heterogeneity models.

Model I. An individual is affected if the individual is a recessive homozygote at *all* the loci involved in the pathogenesis of the disorder. The loci are assumed to be autosomal, unlinked and biallelic. For example, if the disorder is caused by the action of L unlinked loci and at each locus there are two alleles – A,a; B,b; C,c; etc. (a,b,c, ... denoting the recessive alleles) – affected individuals are of genotype aabbcc ...; individuals of all other genotypes are phenotypically normal. Thus, of the 3^L genotypes, only 1 genotype gives rise to the affected phenotype; individuals of the remaining 3^L-1 genotypes are phenotypically normal.

Model II. An individual is affected if the individual is a recessive homozygote at *any one* of the L loci involved. In this case, of the 3^L genotypes, 3^L-2^L genotypes lead to the affected phenotype; the remaining 2^L genotypes lead to the normal phenotype. Thus, for $L=2$, individuals of genotypes AAbb, Aabb, aaBB, aaBb, and aabb are phenotypically affected, and those of genotypes AABB, AABb, AaBB, AaBb are phenotypically normal.

Population prevalence

Consider a disorder determined by the epistatic action of recessive alleles at multiple unlinked loci (Model I). Suppose, q_i denotes the frequency of the recessive allele at the biallelic locus $i\,(i=1,2,\ldots,L)$ in a population. If the population practises random mating, then the prevalence (δ) of the disorder in the population is:

$$\delta=\prod_{i=1}^{L} q_i^2 .$$

Table 1. Two-locus genotypes and their frequencies in a population in Hardy–Weinberg equilibrium

		Locus 2		
Loci		BB	Bb	bb
Locus 1	AA	AABB ($p_1^2p_2^2$)	AABb ($2p_1^2p_2q_2$)	AAbb ($p_1^2q_2^2$)
	Aa	AaBB ($2p_1q_1p_2^2$)	AaBb ($4p_1q_1p_2q_2$)	Aabb ($2p_1q_1q_2^2$)
	aa	aaBB ($q_1^2p_2^2$)	aaBb ($2q_1^2p_2q_2$)	aabb ($q_1^2q_2^2$)

Table 2. Penetrances of two-locus genotypes

		Locus 2		
Loci		BB	Bb	bb
Locus 1	AA	h_1	h_2	h_3
	Aa	h_4	h_5	h_6
	aa	h_7	h_8	h_9

If $q_i = q$ (for all $i = 1, 2, 3, \ldots, L$), then,

$$\delta = q^{2L}, \qquad [0 < \delta,\ q < 1].$$

For a disorder which is due to recessive homozygosity at any one of the L loci involved (Model II)

$$\delta = \left[1 - \prod_{i=1}^{L} (1 - q_i^2)\right].$$

If $q_i = q$ (for all $i = 1, 2, \ldots, L$), then,

$$\delta = [1 - (1 - q^2)^L], \qquad [0 < \delta,\ q < 1].$$

For Model I, the prevalence, for a fixed value of the allele frequency q, decreases sharply with the increase in the number of loci L. For Model II, however, the prevalence increases with the increase of the number of loci. The prevalence, for a fixed value of L, increases with increase in the allele frequency q for both models. We also note that for both models the prevalence is exactly same when $L = 1$.

For a recessive disorder, it is of interest to examine its characteristics in an inbred population. Suppose, a population practises inbreeding and the inbreeding coefficient is $F(F > 0)$. Then, at the ith locus ($i = 1, 2, 3, \ldots, L$), the frequency of the recessive homozygote will be $q_i F + q_i^2(1 - F)$ for Model I and $1 - [1 - \{q_i F + q_i^2(1 - F)\}]$ for Model II. If $q_i = q$ ($i = 1, 2, 3, \ldots, L$), then, prevalence of a disorder for Models I and II will be:

$$\delta = [qF + q^2(1 - F)]^L$$

and,

$$\delta = 1 - [1 - \{qF + q^2(1 - F)\}]^L.$$

As is expected, for fixed values of q and L, prevalence increases with increase in the population inbreeding coefficient F. Further, the rate of decrease in population prevalence under Model I, or the rate of increase in population prevalence under Model II, with increase in L for a fixed value of q is dependent on the value of the population inbreeding coefficient, F.

Methodology for analysis of family data

Data pertaining to phenotypes, ages, etc. on members of families are statistically analysed for parsimonious model selection. Under each plausible model, likelihood of the data on a set of families is evaluated. As has been mentioned earlier, families are generally ascertained through the presence of an affected individual, who is called the proband. Such non-random ascertainment of families necessitates appropriate corrections in the likelihood function. In this section, we shall derive the likelihood functions of data on various types families under different non-random ascertainment schemes.

Likelihood of a normal × normal family ascertained through an affected offspring

To determine the likelihood of a normal × normal family ascertained through an affected offspring, we first note that each of the normal parents must either be heterozygous at each of the L loci or recessive homozygous at all L loci except at least one. This is because, to produce an affected (aabbcc . . .) offspring, each parent must be capable of transmitting an abc . . . gamete, and the reason why neither parent can be recessive homozygous at all the L loci is that each parent is known to be phenotypically normal. Hence, for any such family in which the father is heterozygous at i loci and the mother is heterozygous at j loci, the mating frequency, M_{ij} is:

$$M_{ij} = \frac{\binom{L}{i}\binom{L}{j} H^{i+j} R^{2L-i-j}}{\sum_{i=1}^{L}\sum_{j=1}^{L}\binom{L}{i}\binom{L}{j} H^{i+j} R^{2L-i-j}}.$$

The probability θ_{ij} that this family produces an affected offspring is:

$$\theta_{ij} = \frac{1}{2^{i+j}}.$$

Now, the probability, α_r, that a family with r affected offspring will have at least one proband is[38]:

$$\alpha_r = 1 - (1 - \pi)^r,$$

where π denotes the conditional probability that an offspring is a proband given that (s)he is affected, which we assume is independent of the parental mating type.

The probability, $\tau_{sr}^{(ij)}$ ($r = 1, 2, \ldots, s$), that a family of ijth type (that is, in which the father is heterozygous at i loci and the mother is heterozygous at j loci) of size s will have r affected offspring is:

$$\tau_{sr}^{(ij)} = \binom{s}{r} \theta_{ij}^r (1 - \theta_{ij})^{s-r}.$$

Therefore, the probability, $\varphi_{sr}^{(ij)}$, that a family of ijth type of size s will have r affected offspring and will be ascertained is:

$$\varphi_{sr}^{(ij)} = \tau_{sr}^{(ij)} \cdot \alpha_r; \quad r = 1, 2, 3, \ldots, s.$$

Hence, the probability, $\Phi_s^{(ij)}$, of a family of ijth type of size s having at least one affected child and being ascertained is:

$$\Phi_s^{(ij)} = \sum_{r=1}^{s} \varphi_{sr}^{(ij)} = 1 - (1 - \pi\theta_{ij})^s$$

It, therefore, follows that the probability that a family of size s will have at least one affected offspring and will be ascertained is:

$$\sum_{i=1}^{L}\sum_{j=1}^{L} M_{ij}\Phi_s^{(ij)} = \sum_{i=1}^{L}\sum_{j=1}^{i} M_{ij}[1 - (1 - \pi\theta_{ij})^s].$$

The likelihood, $\mathcal{L}$, of an ascertained family of size s having r affected offspring is:

$$\mathcal{L} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{L} M_{ij}\varphi_{sr}^{(ij)}}{\sum_{i=1}^{L}\sum_{j=1}^{L} M_{ij}\Phi_s^{(ij)}}.$$

Under single ascertainment, $\alpha_r = r\pi$. Hence, $\mathcal{L}$ reduces to:

$$\mathcal{L} = \binom{s}{r}\left(\frac{r}{s}\right) \times$$

$$\frac{\sum_{i=1}^{L}\sum_{j=1}^{L} \binom{L}{i}\binom{L}{j} 2^{(i+j)(1-\delta)} (2^{i+j} - 1)^{s-r} p^{i+j} q^{4L-i-j}}{\sum_{i=1}^{L}\sum_{j=1}^{i} \binom{L}{i}\binom{L}{j} p^{i+j} q^{4L-i-j}},$$

where $p = 1 - q$.

Although the above equation looks formidable, it can be considerably simplified because several mating types have the same values of i and j, and consequently the same value of θ_{ij}. This is exemplified in Table 3.

While the above likelihood equation has been derived for unrelated parents, extension to the situation when the parents are related is straightforward. The likelihood function remains valid; the only modification that is necessary is in the mating probabilities. These changed probabilities can be derived using the I-T-O method[39].

For the two-locus model, the unconditional mating probabilities as given in Table 3 for unrelated parents change to:

when parents are an uncle–niece pair:

class 1: $p^2q^2(1/2 + 2pq)^2$;

class 2: $p^2q^3(1 + 4pq)(1 + 2q)$;

class 3: $2pq^4\{[(1 + q)(1 + 4q) + p(1 + 2q)^2]/4\}$;

when parents are a pair of first cousins:

class 1: $1/16 + 3p^3q^3(1/2 + 3pq)$;

class 2: $p^2q^3[1/4 + 3q/2 + 3pq(1 + 6q)]$;

class 3: $pq^4[(1 + q) + 12pq(1 + 3q)]/4$.

Incorporation of variable age at onset

Preliminaries and notations. Under the multilocus models considered, an individual of a given phenotype may potentially be of any one of several genotypes. For example, under Model I, a phenotypically normal individual can be of any one of 8 genotypes (AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb) if two loci are considered, while under Model II, such an individual can be of any one of 4 genotypes (AABB, AABb, AaBB, AaBb). Although under Model I, an affected individual is necessarily of genotype aabb if two loci are involved, under Model II such an individual can be of any one of 5 genotypes (aaBB, aaBb, aabb, AAbb, Aabb). Late age at onset adds to the list of potential genotypes of normal individuals. For example, under Model I, a normal individual may also be of genotype aabb but may not have manifested the disorder at the age at examination.

When two loci are involved, we present in Table 4, the list of various possible genotypic matings, mating probabilities, phenotypic mating types and segregation probabilities, separately for Models I and II. While listing the phenotypic mating types in this table, the possibility that an individual may be of the susceptible

Table 3. Parental genotypic mating classes, segregation probabilities and mating frequencies for Model I with $L = 2$

Class	Genotypic mating Father × Mother	No. of heterozygous loci: Father (i)	No. of heterozygous loci: Mother (j)	Probability of affected offspring (θ_{ij})	Mating probability for class: Unconditional	Mating probability for class: Conditional
1	AaBb × AaBb	2	2	1/16	$16p^4q^4$	p^2
2	AaBb × aaBb	2	1	1/8	$32p^3q^5$	$2pq$
	aaBb × AaBb	1	2	1/8		
	AaBb × Aabb	2	1	1/8		
	Aabb × AaBb	1	2	1/8		
3	aaBb × aaBb	1	1	1/4	$16p^2q^6$	q^2
	aaBb × aaBb	1	1	1/4		
	Aabb × aaBb	1	1	1/4		
	Aabb × Aabb	1	1	1/4		

genotype (e.g., aabb under Model I) but may not have manifested the disorder because of late age at onset has not been taken into account. This possibility introduces a complexity. For example, under Model I, when variable age at onset is considered, the genotypic mating

Table 4. Genotypic and phenotypic mating types, their probabilities and segregation probabilities when two loci are involved

Sl. no.	Genotypic mating type	Mating prob.	Segregation prob. Model I	Segregation prob. Model II	Phenotypic mating type Model I	Phenotypic mating type Model II
1	AABB × AABB	p^8	0	0	Nor × Nor	Nor × Nor
2	AABB × AABb	$4p^7q$	0	0	Nor × Nor	Nor × Nor
3	AABB × AAbb	$2p^6q^2$	0	0	Nor × Nor	Nor × Aff
4	AABB × AaBB	$4p^7q$	0	0	Nor × Nor	Nor × Nor
5	AABB × AaBb	$8p^6q^2$	0	0	Nor × Nor	Nor × Nor
6	AABB × Aabb	$4p^5q^3$	0	0	Nor × Nor	Nor × Aff
7	AABB × aaBB	$2p^6q^2$	0	0	Nor × Nor	Nor × Aff
8	AABB × aaBb	$4p^5q^3$	0	0	Nor × Nor	Nor × Aff
9	AABB × aabb	$2p^4q^4$	0	0	Nor × Aff	Nor × Aff
10	AABb × AABb	$4p^6q^2$	0	1/4	Nor × Nor	Nor × Nor
11	AABb × AAbb	$4p^5q^3$	0	1/2	Nor × Nor	Nor × Aff
12	AABb × AaBB	$8p^6q^2$	0	0	Nor × Nor	Nor × Nor
13	AABb × AaBb	$16p^5q^3$	0	1/4	Nor × Nor	Nor × Nor
14	AABb × Aabb	$8p^4q^4$	0	1/2	Nor × Nor	Nor × Aff
15	AABb × aaBB	$4p^5q^3$	0	0	Nor × Nor	Nor × Aff
16	AABb × aaBb	$8p^4q^4$	0	1/4	Nor × Nor	Nor × Aff
17	AABb × aabb	$4p^3q^5$	0	1/2	Nor × Aff	Nor × Aff
18	AAbb × AAbb	p^4q^4	0	1	Nor × Nor	Aff × Aff
19	AAbb × AaBB	$4p^5q^3$	0	1/2	Nor × Nor	Aff × Nor
20	AAbb × AaBb	$8p^4q^4$	0	1/2	Nor × Nor	Aff × Nor
21	AAbb × Aabb	$4p^3q^5$	0	1	Nor × Nor	Aff × Aff
22	AAbb × aaBB	$2p^4q^4$	0	0	Nor × Nor	Aff × Aff
23	AAbb × aaBb	$4p^3q^5$	0	1/2	Nor × Nor	Aff × Aff
24	AAbb × aabb	$2p^2q^6$	0	1	Nor × Aff	Aff × Aff
25	AaBB × AaBB	$4p^6q^2$	0	1/4	Nor × Nor	Nor × Nor
26	AaBB × AaBb	$16p^5q^3$	0	1/4	Nor × Nor	Nor × Nor
27	AaBB × Aabb	$8p^4q^4$	0	1/4	Nor × Nor	Nor × Aff
28	AaBB × aaBB	$4p^5q^3$	0	1/2	Nor × Nor	Nor × Aff
29	AaBB × aaBb	$8p^4q^4$	0	1/2	Nor × Nor	Nor × Aff
30	AaBB × aabb	$4p^3q^5$	0	1/2	Nor × Aff	Nor × Aff
31	AaBb × AaBb	$16p^4q^4$	1/16	7/16	Nor × Nor	Nor × Nor
32	AaBb × Aabb	$16p^3q^5$	1/8	1/2	Nor × Nor	Nor × Aff
33	AaBb × aaBB	$8p^4q^4$	0	1/2	Nor × Nor	Nor × Aff
34	AaBb × aaBb	$16p^3q^5$	1/8	5/8	Nor × Nor	Nor × Aff
35	AaBb × aabb	$8p^2q^6$	1/4	3/4	Nor × Aff	Nor × Aff
36	Aabb × Aabb	$4p^2q^6$	1/4	1	Nor × Nor	Aff × Aff
37	Aabb × aaBB	$4p^3q^5$	0	1/2	Nor × Nor	Aff × Aff
38	Aabb × aaBb	$8p^2q^6$	1/4	3/4	Nor × Nor	Aff × Aff
39	Aabb × aabb	$4pq^7$	1/2	1	Nor × Aff	Aff × Aff
40	aaBB × aaBB	p^4q^4	0	1	Nor × Nor	Aff × Aff
41	aaBB × aaBb	$4p^3q^5$	0	1	Nor × Nor	Aff × Aff
42	aaBB × aabb	$2p^2q^6$	0	1	Nor × Aff	Aff × Aff
43	aaBb × aaBb	$4p^2q^6$	1/4	1	Nor × Nor	Aff × Aff
44	aaBb × aabb	$4pq^7$	1/2	1	Nor × Aff	Aff × Aff
45	aabb × aabb	q^8	1	1	Aff × Aff	Aff × Aff

AABB × aabb may phenotypically either be normal × affected or normal × normal.

For the formulation of likelihood of phenotypic observations on offspring given the parental phenotypic mating type, the following further preliminaries and notations are in order.

1. For a particular phenotypic mating type, several genotypic mating types are possible. If the mating involves parent(s) who is (are) phenotypically normal, then the current age(s) of the parent(s) also need to be taken into consideration while enumerating the possible genotypic matings. We shall denote as g_f and g_m, the current ages of father and mother respectively.

2. We shall denote as: z_x = Prob {an individual of age x is phenotypically normal given that (s)he is of the susceptible genotype(s)}. These probabilities are estimated from age at onset data of affected individuals. In practice, it may be necessary to form age-groups to avoid vagaries of small sample sizes. When age groups are formed, z_i will denote the above conditional probability for an individual belonging to age-group i; $i = 1, 2, \ldots, G$.

3. We shall denote as: μ_k = Prob {genotypic mating type is k given the phenotypic mating type and age(s) of the phenotypically normal parent(s)}; $k = 1, 2, \ldots, K$ = number of genotypic matings for a specified phenotypic mating. These are calculated straightforwardly from the mating probabilities given in Table 4. However, these probabilities need to be multiplied by appropriate z_i values in specific cases. For example, under Model I, given a normal × affected mating, K should equal 8 (corresponding to serial numbers 9, 17, 24, 30, 35, 39, 42 and 44 of Table 4) if the disorder expresses itself at birth. However, when the disorder has a late age at onset, a normal × affected mating may also be of type aabb × aabb (serial number 45 of Table 4. Thus, $K = 9$. The mating probability of the aabb × aabb mating given that the phenotypic mating type is normal × affected, and that the normal individual belongs to ith age group is $q^8 z_i$. The conditional probabilities, μ_k's, are obtained by dividing the unconditional probabilities by the sum of the unconditional probabilities of all genotypic matings corresponding to the given phenotypic mating.

4. We shall denote as: θ_k = Prob {offspring is of a susceptible genotype given that the parental genotypic mating is of type k}. For example, under Model I, θ_k = Prob {offspring is of genotype aabbcc ... given that the parental genotypic mating is of type k}. But under Model II, θ_k = Prob {offspring is of AAbb or Aabb or aaBB or aaBb or aabb given the parental genotypic mating of type k}. These are also given in Table 4.

5. Consider an offspring of age x in a family in which parental genotypic mating is of type k. The probability of this offspring being phenotypically affected

is $\theta_k(1-z_x)$, and of being phenotypically normal is $1-\theta_k+\theta_k z_x = 1-\theta_k(1-z_x)$.

6. For a particular nuclear family, we shall denote as: n_i = total number of offspring in age-group i; m_i = number of affected offspring in age-group i; $(n_i - m_i)$ = number of normal offspring in age-group i.

Likelihood function for a normal × affected family, ascertained through an affected parent

The data comprise numbers of affected offspring belonging to each of the G age groups; that is, m_i and $n_i - m_i$; $i = 1, 2, \ldots, G$. Given that parents are normal × affected, one can enumerate all possible genotypic matings that can give rise to a normal × affected phenotypic mating, under either Model I or Model II. Suppose K such genotypic matings are possible. For each genotypic mating, k, the conditional mating probability μ_k can be worked out as indicated in the previous section after taking into account the age of the normal parent. For a given genotypic mating, k, the likelihood of phenotypic observations of offspring belonging to age group i is:

$$\binom{n_i}{m_i}[\theta_k(1-z_i)]^{m_i}\,[1-\theta_k(1-z_i)]^{n_i-m_i}.$$

Thus, the conditional likelihood function of phenotypic observations on all offspring given the parental mating type is:

$$\mathcal{L} = \sum_{k=1}^{K} \mu_k \prod_{i=1}^{G} \binom{n_i}{m_i} [\theta_k(1-z_i)]^{m_i}\,[1-\theta_k(1-z_i)]^{n_i-m_i}.$$

Likelihood function for a normal × normal family, ascertained through an affected offspring

In comparison with the previous case, a normal × affected family ascertained through an affected offspring raises two problems. First, the ages of both normal parents have to be considered in determining μ_k's. For example, under Model I, for $L = 2$, a normal × normal mating may actually be of type aabb × aabb. That is, both parents can be of the susceptible genotype (aabb), without manifesting the disorder at the time of data collection. The unconditional probability of this genotypic mating will be $q^8 z_i z_j$, when the parents belong to age groups i and j $(i, j = 1, 2, \ldots, G)$. Second, while no correction for bias of ascertainment was required in the previous case (normal × affected family ascertained through an affected parent), when a family is ascertained through an affected offspring, the likelihood has to be corrected for ascertainment-bias. The likelihood function, $\mathcal{L}$, can be written as:

$$\mathcal{L} = [\alpha_m \cdot \ell(\mathbf{n}, \mathbf{m})]/\beta(\mathbf{n}, \mathbf{m}),$$

where the form of the function $\ell(\mathbf{n}, \mathbf{m})$; $\mathbf{n}$ and $\mathbf{m}$ being vectors, is the same as the likelihood function of the previous case. [Of course, enumeration of genotypic matings and calculation of conditional mating probabilities will correspond to a normal × normal phenotypic mating rather than a normal × affected mating.] If $m = \sum_{i=1}^{G} m_i$ denotes the total number of affected offspring in the family, then α_m = Prob (a family with r affected offspring will have at least one proband) $= 1-(1-\pi)^m$, where π denotes the probability of ascertainment. Thus, the numerator of $\mathcal{L}$, $\alpha_m \cdot \ell(\mathbf{n}, \mathbf{m})$, denotes the likelihood that in a family with n_i offspring there will be m_i affected in age group $i (i = 1, 2, \ldots, G)$, and that such a family will be ascertained. The denominator of $\mathcal{L}$, $\beta(\mathbf{n}, \mathbf{m})$, denotes the probability that a family with n_i offspring in age group i has at least one affected offspring and is ascertained. This term is obtained as:

$$\beta(\mathbf{n}, \mathbf{m}) = \sum_{r=1}^{N} \left[\alpha_r \cdot \sum_{\substack{\ell = (\ell_1, \ell_2, \ldots, \ell_G) \\ \ell_i \le n_i \\ \sum \ell_i = r}} l(\mathbf{n}, \mathbf{l}) \right],$$

where

$$N = \sum_{i=1}^{G} n_i.$$

When, $\pi \approx 0$, the likelihood function simplifies to:

$$\mathcal{L} = r \cdot \ell(\mathbf{n}, \mathbf{m})/\beta(\mathbf{n}, \mathbf{m}),$$

where

$$\beta(\mathbf{n}, \mathbf{m}) = \sum_{r=1}^{N} r \cdot \sum \ell(\mathbf{n}, \mathbf{l}),$$

and the range and constraints of the second summation are those of $\beta(\mathbf{n}, \mathbf{m})$ given earlier.

Computations of likelihood functions: Some comments

The number of possible genotypic matings, K, for a given phenotypic mating type increases drastically with increase in the number of loci, L. For a fixed value of L, K is also much larger if a disorder has a variable onset age compared to one which is expressed at birth. Thus, for a disorder with a variable onset age, the

number of terms to be summed in the likelihood function is usually large. However, several genotypic matings have the same segregation probability, as is evident from Table 4. Considerable computational simplification is obtained by pooling all genotypic matings with the same segregation probability. When this is done, the number of terms to be summed in the likelihood function reduces to the number of distinct values of the segregation probability.

When data on a number of nuclear families of a specific mating type are available, the joint likelihood is the product of likelihoods of individual families. Here again, considerable computational simplification is obtained by pooling data of all families in which the normal parent(s) belongs to the same age group(s).

Discussion

Starting with an overview of the development of multilocus models, we have focused on two models for complex human disorders – epistatic and heterogeneity models. We have provided the methodology for calculating the likelihood of data on families for each of these models. Because we have considered the practical situations that the disorder may have a variable age at onset and that the families from which data are collected may be ascertained through affected individuals, the proposed methodology should be widely applicable. Based on the value of the likelihood function of data on a set of families, model selection can easily be made. Although, in this paper, we have provided no application of the proposed methodology, this can be found in Nath *et al.*[35] and Nath[36].

1. Lander, E. S. and Schork, N. J., *Science*, 1994, **265**, 2037–2048.
2. Majumder, P. P., in *Human Population Genetics: A Centennial Tribute to J.B.S. Haldane* (ed. Majumder, P. P.), Plenum Press, New York, 1993, pp. 89–98.
3. Vieland, V. J., Hodge, S. E. and Greenberg, D. A., *Genet. Epidemiol.*, 1992, **9**, 45–59.
4. Majumder, P. P., *Am. J. Hum. Genet.*, 1989, **45**, 412–423.
5. Rice, J. P., Neuman, R. J., Burroughs, T. E., Hampe, C. L., Daw, E. W. and Suarez, B. K. *Am. J. Hum. Genet.*, 1993, **53**, A66.
6. Dizier, M-H., Bonaiti-Pellie, C. and Clerget-Darpoux, F., *Am. J. Hum. Genet.*, 1993, **53**, A793.
7. Goldin, L. R. and Weeks, D. E., *Am. J. Hum. Genet.*, 1993, **53**, A1006.
8. Neuman, R. J. and Rice, J. P., *Genet. Epidemiol.*, 1992, **9**, 347–365.
9. Schork, N. J., Boehnke, K., Terwilliger, J. D. and Ott, J., *Am. J. Hum. Genet.*, 1993, **53**, 1127–1136.
10. Tiwari, J. L. and Terasaki, P. I., *HLA and Disease Associations*, Springer, Berlin, 1985.
11. Thomson, G., *Ann. Hum. Genet.*, 1980, **43**, 383–398.
12. Wendt, G. G. and Drohm, D., *Fortschritte der Allgemeinen und Klineschen Humangenetik. Vol IV: Die Huntingtonsche Chorea*, Stuttgart, Thieme, 1972.
13. Haldane, J. B. S., *J. Genet.*, 1941, **41**, 149–157.
14. Risch, N., *Am. J. Hum. Genet.*, 1990, **46**, 222–228.
15. Riordan, J. R., Rammens, J. M. and Kerem, B., *Science*, 1989, **245**, 1066–1073.
16. Audrezet, M. P., Novelli, G., Mereier, B., Sangiuolo, F., Maceratesi, P., Ferec, C. and Dallapiccola, B., *Hum. Hered.*, 1993, 43, 295–300.
17. Riccardi, V. M. and Eichner, J. E., *Neurofibromatosis: Phenotype, Natural History and Pathogenesis*, Johns Hopkins University Press, Baltimore, 1986.
18. Barker, D., Wright, E., Nguylen, K., Cannon, L., Fain, P., Goldin, D., Bishop, D. T., Carey, J., Baty, B., Kivlin, J., Willard, H., Waye, J. S., Greig, G., Leinward, L., Nakamura, Y., O'Connell, P., Leppert, M., Lalouel, J. M., White, R. and Skolnick, M., *Science*, 1987, **236**, 1100–1102.
19. Seizinger, B., Ozelius, L. J., Lane, A. H., George-Hyslop, P., Huson, S., Gusella, J. F. and Martuza, R. L., *Science*, 1987, **136**, 317–319.
20. Majumder, P. P., Ramesh, A., Chinnappan, D., *Am. J. Hum. Genet.*, 1989, **44**, 86–99.
21. Slack, J. and Evans, K. A., *J. Med. Genet.*, 1966, **3**, 239–257.
22. Goldbourt, U. and Neufeld, H. N., *Arteriosclerosis*, 1986, 6, 357–377.
23. Emery, A. E. H., *Methodology in Medical Genetics*, Churchill-Livingstone, Edinburgh, 1976.
24. Passarge, E., *Nature Genet.*, 1993, **4**, 325–326.
25. Angrist, M., Kauffman, E., Slaugenhaupt, S. A., Matise, T. C., Puffenberger, E. G., Washington, S. S., Lipson, A., Cass, D. T., Reyna, T., Weeks, D. E., Sieber, W. and Chakravarti, A., *Nature Genet.*, 1993, **4**, 351–356.
26. Chakravarti, A., in *Genetics of Cellular, Individual, Family, and Population Variability* (eds Sing, C. F. and Hanis, C. L.), Oxford University Press, New York, 1993, pp. 131–136.
27. Steinberg, A. G., Becker, S. W., Fitzpatrick, T. B. and Kierland, R. R., *Am. J. Hum. Genet.*, 1951, **3**, 267–281.
28. Li, C. C., *Am. J. Hum. Genet.*, 1953, **5**, 269–279.
29. Li, C. C., *Am. J. Hum. Genet.*, 1987, **41**, 517–523.
30. Merry, A., Roger, J. H. and Curnow, R. N., *Ann. Hum. Genet.*, 1979, **43**, 71–80.
31. Greenberg, D. A., *Am. J. Hum. Genet.*, 1981, **33**, 519–530.
32. Greenberg, D. A. and Lange, K. L., *Am. J. Med. Genet.*, 1982, **12**, 75–82.
33. Majumder, P. P., Das, S. K. and Li, C. C., *Am. J. Hum. Genet.*, 1988, **43**, 119–125.
34. Majumder, P. P. and Nath, S. K., *J. Genet.*, 1992, **71**, 89–103.
35. Nath, S. K., Majumder, P. P. and Nordlund, J. J., *Am. J. Hum. Genet.*, 1994, **55**, 981–990.
36. Nath, S. K., Ph D thesis (unpublished), Indian Statistical Institute, Calcutta, 1994.
37. Hartl, D. L. and Maruyama, T., *J. Theor. Biol.*, 1968, **20**, 129–163.
38. Elandt-Johnson, R. C., *Probability Models and Statistical Methods in Genetics*, John Wiley & Sons, New York, 1971.
39. Li, C. C. and Sacks, L., *Biometrics*, 1954, **10**, 347–360.