

A Simple Method for Testing Two-Locus Models of Inheritance

DAVID A. GREENBERG¹

SUMMARY

A graphic method for testing simple two-locus models of inheritance is developed. The model assumes two alleles at each locus where both loci exhibit dominant, both exhibit recessive, or one locus exhibits dominant and one locus exhibits recessive inheritance. Examples of applying the graphs using published data on three diseases are given.

INTRODUCTION

The number of models of the mode of inheritance that can easily be tested by mathematical and clinical investigators is limited. Oligo- and polygenic and multifactorial models are often beyond the capability of all but the computerwise and are often not excludable with the amount of data (e.g., number of pedigrees) available. It is also not uncommon to see the concept of reduced penetrance invoked as an explanation for the failure of a segregation ratio calculation to yield a segregation ratio on the order of .25 or .5. While the concept of reduced penetrance (i.e., where individuals possessing the disease genotype may not exhibit the trait or disease) is a useful one and the description is of a real phenomenon, it is probably too often used to cloak our ignorance of the genetics of a disease rather than to display our knowledge.

Recently, there have been reports of human disease that are possibly the result of the interaction of two loci, either linked or unlinked [1, 2]. (The mathematics of two *linked* loci have also been examined recently [3].) In what is apparently a little-known paper, Defrise-Gussenhoven [4] in 1962 proposed using graphs of Snyder's ratio and population prevalence as a way of testing whether such data were consistent with two-locus models of inheritance. (Snyder's ratio is the segregation

Received August 25, 1980; revised November 3, 1980.

This study was supported in part by grant AM-17328 from the National Institutes of Health.

¹ Division of Medical Genetics, Department of Pediatrics, Harbor-UCLA Medical Center, UCLA School of Medicine, Torrance, CA 90509.

© 1981 by the American Society of Human Genetics. 0002-9297/81/3304-0003\$02.00

ratio conditioned on parental phenotype.) However, data from the literature seldom include parental phenotype, data which are necessary for the calculation of Snyder's ratio. Here, graphs similar to those published by Defrise-Gussenhoven are presented, but with the approach of using the population segregation ratio instead of Snyder's ratio. The population segregation ratio can be viewed as Snyder's ratio averaged over all parental phenotypes. While there are 50 possible phenograms that can be produced by two loci with two alleles at each locus [5], only three are considered here. Using the population segregation ratio graphs and prevalence graphs, simple two-locus models of inheritance can easily be tested using data from the literature. Also, several examples of applying the graphs are included to illustrate their utility in discriminating among several models. Two-locus models that are not excluded by this method can then be more specifically tested by using the method described in [4] or by segregation analysis [5].

Here, the graphs presented allow testing of three different two-locus models with two alleles at each locus, one allele "normal" and one trait- (or disease-) producing. The three models are: (1) both loci require a "double dose" of the trait alleles (recessive-recessive or R-R model); (2) one requires only one trait allele and the other requires two (dominant-recessive or D-R model); and (3) the dominant-dominant or D-D model.

METHODS

Three assumptions went into producing the graphs: (1) Hardy-Weinberg equilibrium and random mating prevail; (2) linkage equilibrium between the loci; and (3) penetrance of the trait is close to unity, as reflected by the monozygotic (MZ) twin concordance rate. Penetrance is generally computed by comparing the observed segregation ratio with that predicted by a model, usually simple Mendelian dominant or recessive. Since the present analysis assumes that a second locus is responsible for the "reduced penetrance," it would obviously defeat our purpose to infer reduced penetrance from the failure of a segregation ratio computation to yield .25 or .5. The graphs assume the trait under investigation is either completely genetically determined or, as in the case of coeliac disease (see below), has an environmental component that is ubiquitous.

The population prevalence was taken as the frequency of the appropriate susceptible genotype(s). Table 1 shows the Hardy-Weinberg frequencies for each of the nine possible genotypes, as well as indicating which genotypes are affected under the different models.

The population segregation ratio was calculated according to the formula [6]:

$$\psi = \frac{\sum_i S_i \psi_i}{\sum_i S_i \alpha_i} \quad \alpha_i = \begin{cases} 1, & \text{when } \psi_i > 0 \\ 0, & \text{when } \psi_i = 0 \end{cases}$$

Here ψ_i is the proportion of affected offspring for mating type i and S_i is the frequency of the mating type. This population segregation ratio is the average percentage of affected offspring (from mating types capable of producing affected offspring) weighted by the frequency of those mating types.

The requirements for the data being used to test the two-locus models are stringent. In addition to an MZ twin rate of unity, estimates of the population segregation ratio need to be corrected for ascertainment bias. Obviously, if ascertainment is not corrected, the segregation ratio derived from the data will be higher than in reality. The method described here makes no assumptions about the type of ascertainment bias correction used on the data.

TABLE 1
GENOTYPES AND POPULATION FREQUENCIES FOR THE TWO-LOCUS MODEL

GENOTYPE	HARDY-WEINBERG FORMULA	PHENOTYPE		
		D-D	D-R	R-R
AABB	r^2q^2	+	+	+
AABb	$2r^2q(1 - q)$	+	-	-
AAbb	$r^2(1 - q)^2$	-	-	-
aABB	$2(1 - r)r q^2$	+	+	-
aABb	$2(1 - r)r \cdot 2(1 - q)q$	+	-	-
aAbb	$2(1 - r)r(1 - q)^2$	-	-	-
aaBB	$(1 - r)^2q^2$	-	-	-
aaBb	$(1 - r)^2 \cdot 2(1 - q)q$	-	-	-
aabb	$(1 - r)^2(1 - q)^2$	-	-	-

NOTE: A = trait allele at locus 1, frequency r ; a = all other alleles at locus 1, frequency $(1 - r)$; B = trait allele at locus 2, frequency q ; b = all other locus 2, frequency $(1 - q)$; + = expresses trait; - = does not express trait.

The assumption is also made that the trait or disease being tested is genetically homogeneous. The classification of several distinct diseases as being the same will lead to incorrectly high population prevalences and meaningless segregation ratios. Diseases that have an environmental component will be disqualified from analysis by this method since the penetrance by the MZ twin rate will probably not be one. A disease such as coeliac disease that has a ubiquitous environmental component is not inappropriate to analyze in this way (see below).

A grid was constructed, each point of which had as coordinates the gene frequencies of the trait-producing alleles at the two loci. The assigned value of each such point was then either the population prevalence (figs. 1-3) or segregation ratio (figs. 4-6) defined by the gene frequencies at the point. Lines connecting equal values of the population prevalence (isoprev) or segregation ratio (isoseg) were drawn.

RESULTS

Figures 1 through 6 show the resulting graphs. Figures 4-6 represent the "contour maps" of the segregation ratio for the D-D, D-R, and R-R models, respectively, as a function of the frequency of the trait-producing alleles at each of the two loci. Figures 1-3 show similar graphs for the population prevalence.

To use the graphs, note the area on the segregation ratio map defined by the segregation ratio calculated from the data, plus and minus 1 SD. This region is called the allowed area. Then use a similar procedure with the population prevalence graphs, choosing as the limits the best estimates of population prevalence. If the allowed areas of the population prevalence map and the segregation ratio map do not overlap, then the model is rejected, since the gene frequencies corresponding to the population prevalence are inconsistent with those for the segregation ratio. If the allowed areas do overlap, the model is not rejected.

Note that the simple Mendelian dominant and recessive models are special cases of the more general two-locus models. As the allele frequency at one of the loci

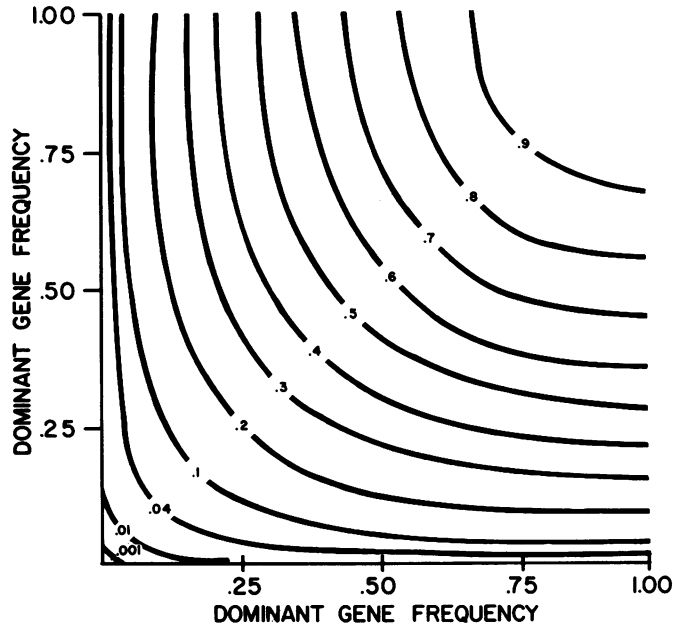


FIG. 1.—Population prevalence contour graph for the D-D model

approaches unity, the segregation ratio will reduce to the simple Mendelian case. If one assumes, for example, the data for Tay-Sachs disease (population prevalence about 1:4,000 in the Jewish population and a segregation ratio of .25) and examines the R-R maps (figs. 3 and 6), one sees that the appropriate isoprev and isoseg lines intersect at a second allele frequency of 1.

The population prevalence, as might be expected, is extremely sensitive to gene frequency, whereas the segregation ratio is much less so. The segregation ratio also has limits for each of the three models below which it is impossible to go, no matter how low the gene frequency. This is not surprising if one remembers that once a gene is segregating in a family, the population frequency becomes irrelevant for the family. Such is not the case with the population prevalence, which can go continuously to zero. As can be seen from figures 1-3, when the gene frequency at one of the loci is high, say, greater than .5, the population prevalence is particularly sensitive to the frequency of the other gene, especially when the frequency of the other gene is relatively low (say, less than .1).

EXAMPLES

To illustrate the use of the graphs, data on three diseases will be examined—coeliac disease, polydactyly, and hemochromatosis. The coeliac disease data will be seen to fit only the R-R model and polydactyly will fit the D-D two-locus model. The data on hemochromatosis represent a more ambiguous situation, with only the R-R model being excludable.

Coeliac Disease

The data for coeliac disease are as follows: Population prevalence is between 1.6×10^{-4} and 6×10^{-4} . The segregation ratio is $.08 \pm .023$ (the data for coeliac disease are discussed more thoroughly in [6]).

The D-D Model. The D-D model is eliminated because the segregation ratio allowed by the model does not go below .25. In addition, a population prevalence of less than 1:1,000 leaves almost no allowed area on the population prevalence map.

The D-R Model. While there is some allowed area on the population prevalence map in the D-R model, the segregation ratio map again excludes this model from consideration. The lowest allowed segregation ratio in the D-R model is .125, while the upper standard error (SE) limit of the observed segregation ratio is only .109.

The R-R Model. The segregation ratio map gives an allowed area in the entire lower left corner of the map, up to about the 0.1 isoseg line. The isoprev lines of .0001 and .0005 bracket the range found for the prevalence of coeliac disease. Combining the two graphs has the effect of "cutting off the tails" of the symmetric allowed area for the population prevalence. Therefore, the R-R model for coeliac disease is not rejected (see fig. 7).

Postaxial Polydactyly (Type B)

Postaxial polydactyly type B is the presence of an extra digit, usually poorly formed. The trait is about 10 times more common in blacks than in Caucasians [7].

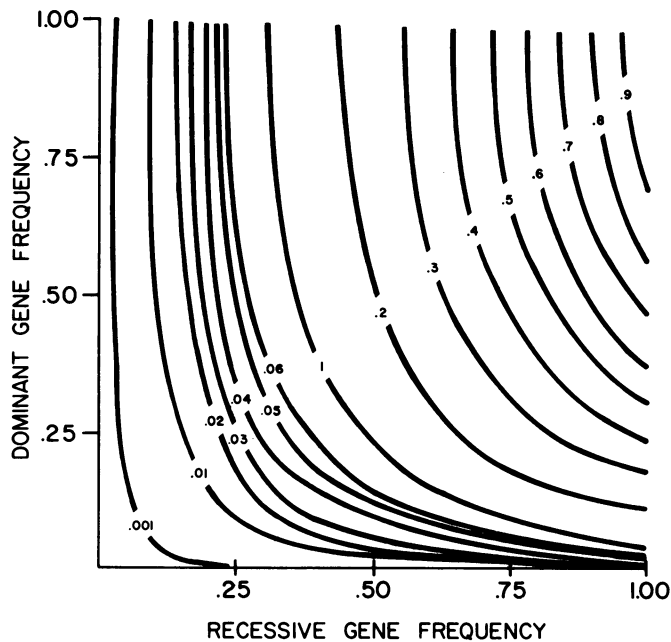


FIG. 2.—Population prevalence contour graph for the D-R model

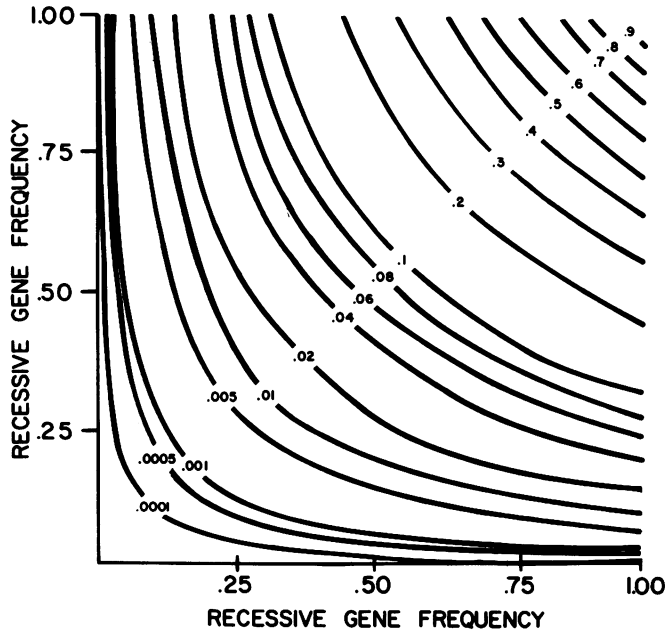


FIG. 3.—Population prevalence contour graph for the R-R model

Walker [8] suggested that the presence of two dominant genes would best explain the pedigree he studied. Scott-Emuakpor and Madeuke [9] found a population prevalence of between about 18 and 27 per 1,000 in Nigeria. The rates differed for males and females, with the female rate being the lower of the two.* These authors also reported a segregation ratio of about .32 with a SE of .02.

The D-D Model. Looking at the population prevalence map (fig. 1), the .01 and .04 isoprev contours bracket the population prevalence estimates. On the segregation ratio map (fig. 4), the bottom left corner from the .3 to the .34 isoseg line is acceptable. Therefore, the overlap or allowed area is mostly limited by the allowed population prevalence area, with the segregation ratio area cutting off the tails and part of the bottom of the symmetric population prevalence area. Therefore, the D-D model is not rejected. The allowed frequencies are from 0 to .25 for each of the loci, with not all combinations being allowed.

D-R Model. Looking at the segregation ratio map, the limits of $.32 \pm .02$ lead to a wide strip of allowed area. However, the overlap area of that strip with the limits of population prevalence leads to an allowed area that is extremely small at the lower end of the segregation ratio limit and at a gene frequency for the dominant gene of almost 1. It is clearly only a very marginal fit and much worse than the D-D model. Therefore, the D-R model is rejected.

R-R Model. The R-R model leads to no allowed area and is rejected.

* There is some ambiguity in the ascertainment bias correction in this study. However, for the purpose of illustrating how to use the graphs, it suffices to use the data as reported. Also, the MZ twin concordance rate is not discussed, so it is assumed to be 1.

Hemochromatosis

While it is not appropriate here to discuss the intricacies of the hemochromatosis story, some explanation is necessary. Hemochromatosis is a disease of iron metabolism characterized by hepatomegaly, and often melanoderma, diabetes, and gonadal insufficiency. Serum iron is elevated, and unsaturated iron-binding capacity is very low.

Hemochromatosis has recently been shown to be a Mendelian recessive disorder with partial expression in heterozygotes [10, 11]. There has been some debate in the past about the mode of inheritance of hemochromatosis due to some complicating factors. For example, females exhibit both the overload and the disease less frequently than males, presumably because of the loss of iron during menses. A further complication is that while the disease tends to aggregate horizontally, there is some vertical aggregation [12]. In addition, the estimates of population prevalence vary from .0001 to .003 [10, 12].

To adhere to the stipulation that the penetrance of the disease be unity, in the following analysis, only data from males are considered, since their penetrance is presumably 1 [12]. Data are taken from Simon et al. [12].

The population segregation ratios are $.26 \pm .057$ when the minor overload is classified as unaffected and $.34 \pm .061$ when the minor overload is called affected. Ascertainment was corrected by the proband method assuming single ascertainment.

We will consider the two segregation ratios under the different assumptions of population prevalence, that is, .0001 and .003.

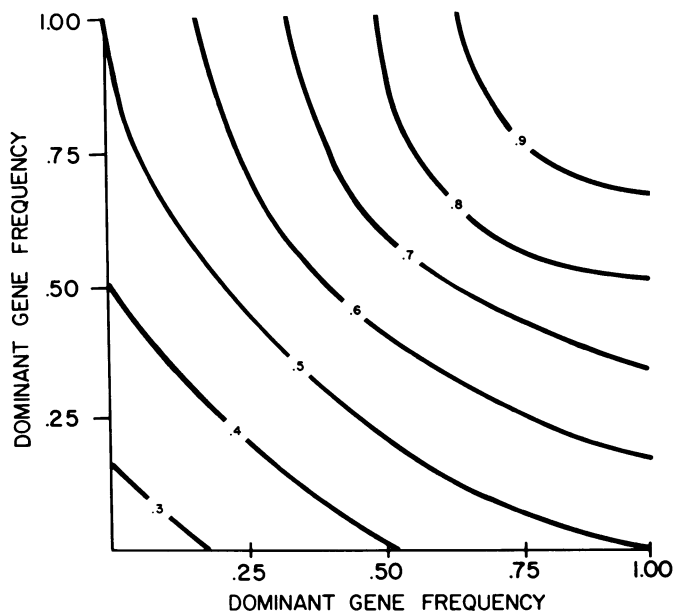


FIG. 4.—Population segregation ratio contour graph for the D-D model. Lowest possible value is .25

D-D Model. If we examine the higher of the two segregation ratios ($.34 \pm .061$), it can be seen that the lowest population prevalence allowed is about 1:100 at minus 1 SE of the segregation ratio. Therefore, if the minor overload is considered affected, the two-locus D-D model does not fit.

Looking at the lower segregation ratio ($.26 \pm .057$), the lower left corner of the segregation ratio map, up to the .3 isoseg, is allowed. Therefore, the limiting factor becomes the population prevalence. A population prevalence of 1:10,000 would require that both gene frequencies be less than .01. The population prevalence in that case becomes very sensitive to the gene frequency. Different gene frequencies that might vary by several percent in different locales would lead to very different prevalences. Therefore, while the D-D model is acceptable under the above conditions, it is less attractive if the population prevalence is on the order of 1:10,000 everywhere. There does exist an allowed area if the population prevalence is on the order of .003, but it is close to the upper end of the observed segregation ratio range. Therefore, if the segregation ratio proves not to be at the upper end of the range, the D-D model could be excluded.

D-R Model. The higher segregation ratio ($.34 \pm .061$) immediately excludes the D-R model from consideration, since predicted population prevalences are too high if that segregation ratio is correct.

When we examine the lower segregation ratio ($.26 \pm .057$), a very narrow (probably negligible) strip of allowed area emerges in the area defined approximately by a dominant gene frequency of between .5 to 1.0 and a recessive gene frequency

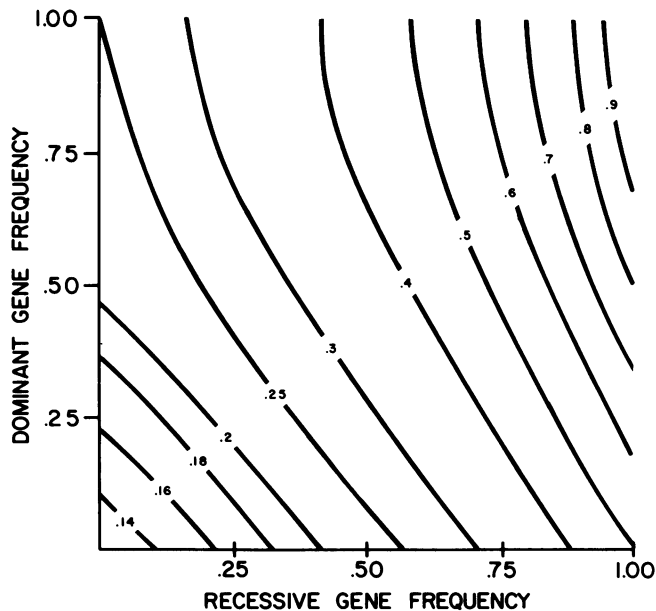


FIG. 5.—Population segregation ratio contour graph for the D-R model. Lowest possible value is .125.

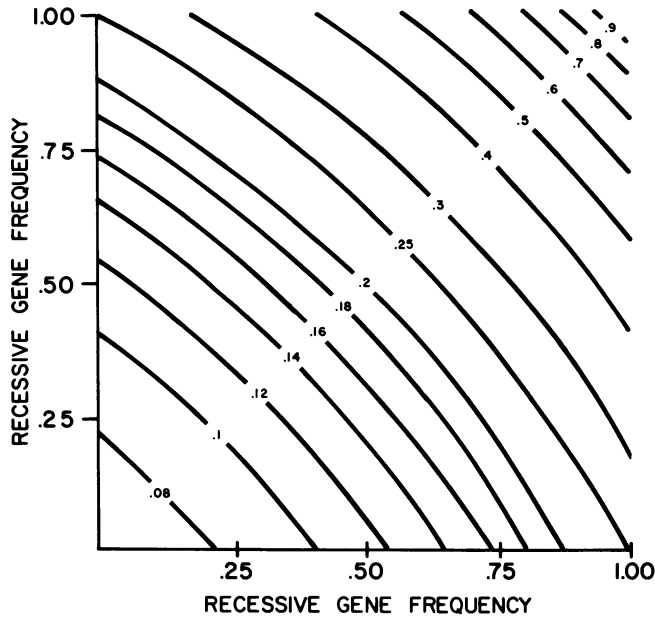


FIG. 6.—Population segregation ratio contour graph for the R-R model. Lowest possible value is .063.

of between 0 and .02 (assuming a population prevalence of 1:10,000). If a population prevalence between .001 and .003 is assumed, the gene frequencies must be .5–1.0 and .05–.1, respectively. These limits illustrate that, while some area is allowed, the frequency of the dominant gene is so high that a simple recessive model (single locus) may be a more attractive hypothesis. (As the trait allele frequency at one of the loci approaches unity, the model reduces to the simple Mendelian case.) If, however, the segregation ratio (still ignoring data from females) is found to be less than .25, the D-R model must be considered. This would mean that the dominant allele is fairly common in the population.

R-R Model. The R-R model can be excluded since the allowed area that appears, assuming the lower segregation ratio, is confined to the areas where the gene frequency of one of the alleles is almost unity (again arguing in favor of a single-locus recessive). The higher segregation ratio leads to no allowed area at all.

To summarize the analysis for hemochromatosis: When the minor overload is classified as affected, all 3 two-locus models can be rejected on the basis of population prevalence. If the minor overload is considered to be unaffected, the R-R model can be excluded, but examination of the R-R graphs tend to support a single-locus recessive. The D-R model does show some allowed area, but only in a portion of the graph where the population prevalence changes rapidly as a function of the recessive gene frequency, again supporting a simple recessive model. Also, if the population prevalence is 1:10,000, the D-R graphs produce almost no allowed area. What area is allowed favors a simple recessive because the frequency of the dominant gene

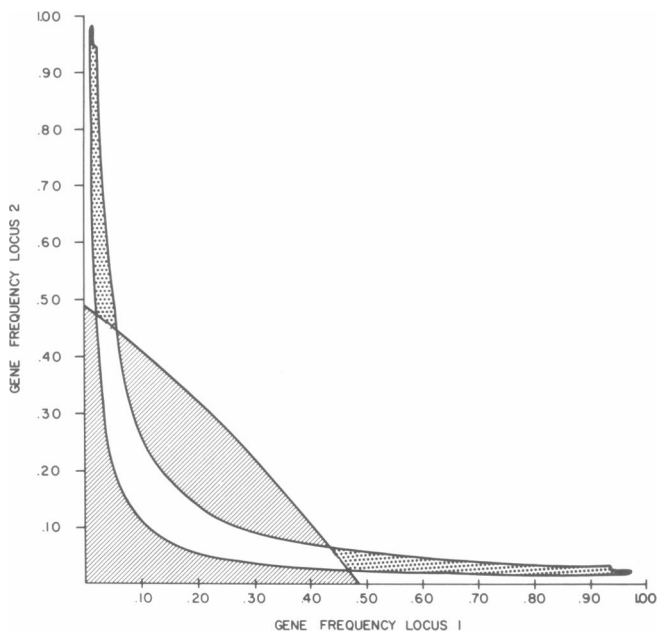


FIG. 7.—Superposition of the allowed areas for coeliac disease. *Stippled area* is the allowed region for the population prevalence; *lined region*, the allowed area for the segregation ratio; and *white area*, the intersection of the two regions.

is high. The D-D model cannot be excluded with the current data. If the lower population prevalence is correct, the allowed area is very small and probably negligible. If the higher population prevalence is a realistic figure, the D-D model is a reasonable one if ψ is greater than .26. More precise definition of the segregation ratio would enable one to choose between simple recessive, D-R, and D-D models.

DISCUSSION

As the above three examples show, it is fairly easy to distinguish among the D-R, D-D, R-R, and simple Mendelian single-locus dominant or recessive models using the population prevalence and segregation ratio graphs. Even in the ambiguous case of hemochromatosis, testing the data against predictions made by different models enables us to identify critical parameters that will distinguish among the models. It must be re-emphasized that these graphs assume a penetrance of 1 for the trait, as indicated by the MZ twin concordance rate. Given a condition such as coeliac disease, where the age of onset may be variable, and that some ambiguities in diagnostic criteria exist, perhaps a concordance rate as low as 70%–80% is still acceptable. A disease such as insulin-dependent juvenile diabetes, however, which appears to be heterogeneous [13] and in which the MZ twin concordance rate may be as low as .2 [14, 15], is inappropriate to analyze in this way without taking account of the heterogeneity (but see [16]).

Obviously, the "testing" method described here 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. If the gene frequencies predicted by the population segregation ratio and trait prevalence are nonoverlapping, the model(s) is obviously not consistent with the observables.

For recessive gene frequencies above about .5, the population prevalence graphs for the R-R and D-R models have an ill-conditioned area when the gene frequency for the second locus is less than about .1. In this region, the population prevalences vary by at least an order of magnitude for a second-locus gene frequency change of only .05. Since this section of the graph is so ill-conditioned, one would hesitate to ascribe a two-locus etiology to a trait or disease if the allowed area fell in this section. If the segregation ratio data, however, excluded a single-locus model, then a two-locus model would have to be considered even if the allowed area fell in the ill-conditioned section. (If the disease were known to have prevalences greatly different in different locales, then such range of gene frequencies would be realistic.)

Note that the models assume Hardy-Weinberg equilibrium and the equal viability of all genotypes (i.e., no selection). Also, the effect of new mutations in the case of the D-R and D-D models has not been taken into account in this analysis.

The objection will undoubtedly be raised that an analysis such as this, or perhaps any mathematical analysis, can really tell us little about the genetic mechanisms of a disease. However, within the rather strict assumptions detailed above, the models presented here can be helpful in two ways: first, they give an explicit and testable meaning to what is loosely called "penetrance" (in this case that another locus is required for disease expression), and second, they indicate what data will eliminate or support a given two-locus model. It is generally accepted that many diseases (cystic fibrosis, Huntington disease) are simple Mendelian recessive or dominant on the basis of the segregation ratio alone.

Examination of the graphs can also give an idea of the effect of even more loci on trait expression. It is clear, for example, that if many loci are involved, the population frequency of some of the trait alleles will probably be high if anything more than sporadic cases are observed.

In summary, the graphs presented here provide a way of testing three different two-locus models of inheritance without the use of a computer. The graphs are relatively easy to use and, in the cases examined, have shown that the criteria of population prevalence and segregation ratio distinguish fairly well among the models. Even in ambiguous cases, the method indicates what data will be able to eliminate some or all the models from consideration.

ACKNOWLEDGMENTS

I wish to express my thanks to Dr. Robert Elston for encouraging this work, and to Dr. Susan Hodge and Dr. Jerome I. Rotter for very helpful discussions.

REFERENCES

1. PENA AS, MANN DL, HAGUE NE, ET AL.: Genetic basis of gluten sensitive enteropathy. *Gastroenterology* 75:230-235, 1978

2. UTERMANN G, VOGELBERG HG, STEINMETZ A, ET AL.: Polymorphism of apolipoprotein E. II. Genetics of hyperlipoproteinemia type III. *Clin Genet* 15:37-62, 1979
3. MERRY A, ROGER JH, CURNOW RN: A two-locus model for the inheritance of familial disease. *Ann Hum Genet* 43:71-80, 1979
4. DEFRISE-GUSSENHOVEN E: Hypothèses de dimérie et de non-pénétrance. *Acta Genet Stat Med (Basel)* 12:65-69, 1962
5. ELSTON RC, RAO DC: Statistical modeling and analysis in human genetics. *Annu Rev Biophys Bioeng* 7:253-286, 1978
6. GREENBERG DA, ROTTER JI: Two locus models for gluten sensitive enteropathy: population genetic considerations. *Am J Med Genet* 8:205-214, 1981
7. FRAZIER TM: A note of race specific congenital malformation rates. *Am J Obstet Gynecol* 80:184-185, 1960
8. WALKER JT: A pedigree of extra-digit polydactyly in a Batutsi family. *Ann Hum Genet* 25:65-68, 1961
9. SCOTT-EMUAKPOR AB, MADEUKE E-DN: The study of genetic variation in Nigeria. II: The genetics of polydactyly. *Hum Hered* 26:198-202, 1976
10. CARTWRIGHT GE, EDWARDS CQ, KRAVITZ K, ET AL.: Hereditary hemochromatosis. *N Engl J Med* 201:175-179, 1979
11. KRAVITZ K, SKOLNICK M, CANNINGS C, ET AL.: Genetic linkage between hereditary hemochromatosis and HLA. *Am J Hum Genet* 31:601-619, 1979
12. SIMON M, ALEXANDER J-L, BOUREL M, LEMAREC B, SCORDIA C: Heredity of idiopathic hemochromatosis: a study of 106 families. *Clin Genet* 11:327-341, 1977
13. ROTTER JI, RIMOIN DL, SAMLOFF IM: Genetic heterogeneity in diabetes mellitus and peptic ulcer, in *Genetic Epidemiology*, edited by MORTON NE, CHUNG DS, New York, Academic Press, 1978, pp 381-414
14. CAHILL GF: Current concepts of diabetic complications with emphasis on hereditary factors: a brief review, in *Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease*, edited by SING CF, SKOLNICK M, New York, Alan R. Liss, pp 113-125
15. GOTTLIEB MS, ROOT HF: Diabetes mellitus in twins. *Diabetes* 17:693-704, 1978
16. THOMSON G: A two locus model for juvenile diabetes. *Ann Hum Genet* 43:383-398, 1980