Acta Genet Med Gemellol 29:255–262 (1980)
The Mendel Institute/Alan R. Liss, Inc.

Received 10 October 1980
Final 24 November 1980

Utility of Genetic Markers in the Study of Human Resemblance

W. J. Kimberling[1,3], D. E. Goldgar[1,2]

[1] *Division of Medical Genetics, Boys Town Institute, Omaha, Nebraska;* [2] *Department of Biometrics, University of Colorado Health Sciences Center, Denver, Colorado;* [3] *Department of Otolaryngology, Creighton University Medical School, Omaha, Nebraska*

A method for the estimation of genetic correlations based upon analysis of genetic marker phenotypes is presented. At a given marker locus, the probability of observing a pair of individuals with a specific combination of phenotypes can be expressed as a function of the gene frequencies at that locus and the genetic correlation (R) between that pair. The likelihood of obtaining a sample of n such pairs with their phenotypes at m marker loci can be expressed as a product of nm such functions. From the likelihood function, maximum likelihood estimates of R can be obtained, and hypotheses about R may be tested. A sample of Swedish twin families (61 dizygotic twin pairs, 268 husband-wife pairs, and 164 sib pairs) were analyzed by this method using information from 21 markers. It was found that for the twin pairs, R = 0.458, which was significantly different from the R calculated for sib pairs (R = 0.558) but not significantly different from the expected 0.5. For the husband-wife pairs, it was found that R = 0.086, which did differ significantly from the expected value of 0, indicating the presence of nonrandom mating in this population.

Key words: Genetic markers, Twin families, Husband-wife correlation, Sib-sib correlation, Dizygotic twins correlation

Genetic correlation is a convenient concept when used to study continuous traits or familial disease which does not have a regular pattern of inheritance. Analytic models for such traits assume certain theoretical correlations between related pairs of individuals. For example, sib pairs are assumed to have genetic correlation of 0.5, half sibs of 0.25, etc. In actual fact, the true sib-sib correlation will vary from one sib pair to the next. Although the potential spread of the correlation is from 0.0 to 1.0, the variance of this correlation is small and narrowly brackets the 0.5 average. Risch and Lange [6] have estimated this variance to be 0.0016, giving approximate 95% confidence limits for the sib-sib correlation of 0.42 and 0.58.

Research conducted at the Departments of Environmental Hygiene of the Karolinska Institute and of the Swedish National Environmental Protection Board. Supported by grant CFTR-1066 from The Council for Tobacco Research – USA, Inc.

0001-5660/80/2904-0255$02.50

The consideration of the variability of sib-sib or other genetic correlations would seem to be unimportant if one assumes that in any given sample, the average genetic correlation of a particular type can be taken to be the corresponding theoretical correlation. However, for reasons such as selection bias or population stratification, etc, this assumption may be unwarranted. In addition, consideration of genetic correlation on an individual pair basis could lead to more powerful genetic analyses of certain quantitative phenotypes [3].

For these reasons we developed a general method of utilizing genetic marker data for the estimation of genetic correlations on both an individual and population basis.

THE METHODS

Swedish Twin Study

In 1976 we became involved in a study of the genetics of smoking behavior using a sample of Swedish twins, their spouses, and offspring. The original purpose of this was to use genetic markers to verify zygosity. Swedish twins were selected on the basis of criteria that have been set forth in a previous publication [1]. Briefly, these are: a) Like-sex pairs; b) ages between 40 and 60; c) both co-twins living; and d) both co-twins married with a living spouse and at least one living child over the age of 20. Twenty-one genetic markers were determined on all 136 twin pairs along with 272 spouses and 358 offspring by standard typing and electrophoretic procedures.

Analysis

There are two phases to the correlation analysis of the Swedish twin data. The first consists of estimating and testing the genetic correlation for each pair of a particular type. Then the results for a particular type (eg, DZ twin pairs) are pooled across all pairs of that type in order to estimate the distribution of the correlations in the population from which the sample was drawn.

From this data the following types of correlations were investigated: the correlation between the DZ twins (TW–TW); the correlation between each twin and his or her spouse (TW–SP); and that between the two spouses who married into the twin family (SP–SP). The offspring of the twins provide three types of correlations, that between full sibs, between half-sibs, and between cousins; however, only the sib-sib correlations were analyzed for this paper. In addition, one spouse from each family was chosen, and these were randomly paired and analyzed so that a baseline distribution could be estimated for the population.

The method proposed here for analyzing the genetic correlations requires the assumption that all of the 21 marker loci are independent. While it is true that some of the marker loci are linked, the effect of this non-independence upon the analyses is believed to be negligible.

The estimate of the genetic correlation between two individuals, expressed in terms of concordance for the marker loci, is derived below.

Given a single locus, A, with an arbitrary number of distinct alleles, there are seven distinguishable genotype combinations for any particular pair of individuals:

I	A_iA_i	A_iA_i
II	A_iA_i	A_iA_j
III	A_iA_i	A_jA_j
IV	A_iA_j	A_iA_j
V	A_iA_i	A_jA_k
VI	A_iA_j	A_iA_k
VII	A_iA_j	A_kA_l

where the indices i, j, k, and l do not represent any specific alleles at locus A, but rather are indicators of how genotypically similar the two members of the pair are. For example, if A is a 3-allele locus with alleles A_1, A_2, A_3, Type I pairs would be A_1A_1–A_1A_1; A_2A_2–A_2A_2, or A_3A_3–A_3A_3; Type VI pairs would be A_1A_2–A_1A_3, A_1A_2–A_2A_3, A_1A_3–A_2A_3, etc.

https://doi.org/10.1017/S0001566000007765 Published online by Cambridge University Press

The probability of observing each of these pair types needs to be expressed as a function of R, the genetic correlation between the pair and p, the vector of allele frequencies at locus A. This probability will be called $F(p, R)$. This function will be derived here only for the pair A_iA_j–A_iA_j.

Be definition the genetic correlation R is the probability (P) of a pair having a gene identical by descent (IBD). Then:

R^2 = P (pair having both alleles IBD)

$2R(1-R)$ = P (pair having exactly one allele IBD)

$(1-R)^2$ = P (pair having no alleles IBD)

If the twins have 2 A_i alleles IBD, then there are only 2 independent A_i genes coming from the population. If the twins share only 1 gene IBD, then there must be 3 independent A_i genes; and if there are no genes IBD, then there have to be 4 independent A_i genes. Thus, the probability of observing an A_iA_j–A_iA_j pair can be expressed as: P (A_iA_j–A_iA_j pair/R, p_i) = P (2 genes IBD and 2 ind A_i alleles) + P (1 gene IBD and 3 ind A_i alleles) + P (0 genes IBD and 4 ind A_i alleles) = $p_i^2 + p_i^3 2R(1-R) + p_i^4(1-R)^2 = p_i^2\ (R^2 + 2\ p_i\ R(1-R) + p_i^2(1-R)^2 = p_i^2\ (R+(1-R)\ p_i)^2$.

The results for the other 7 pair types are similarly derived and we have the following:

	Pair		$F(p,R)$
I	A_iA_i	A_iA_i	$p_i^2\,(R+((1-R)p_i)^2$
II	A_iA_i	A_iA_j	$4\ p_i^2 p_j\ (1-R)\ (R+(1-R)\ p_i)$
III	A_iA_i	A_jA_j	$2\ (1-R)^2\ p_i^2 p_j^2$
IV	A_iA_j	A_iA_j	$2\ p_ip_j\ (R(p_i+p_j)+R^2(1-p_i-p_j)+(1-R)^2\ 2\ p_ip_j)$
V	A_iA_i	A_jA_k	$4\ p_i^2 p_j\ p_k\ (1-R)^2$
VI	A_iA_j	A_iA_k	$4\ p_ip_jp_k\ (1-R)\ (R+2(1-R)\ p_i)$
VII	A_iA_j	A_kA_l	$8\ p_ip_jp_kp_l\ (1-R)^2$

where p is the vector of allele frequencies at locus A.

Substituting R=1, R=½, and R=0 into these functions yields the familiar results for MZ twins, sib pairs, and unrelated individuals, respectively.

These functions hold only for bilineal relationships and unrelated pairs. Special relationships, such as half sibs, require a different formulation. These formulae are sufficient for use here because it is the purpose of this research to test whether full sibs and randomly paired individuals deviate from expectation.

Now that the likelihood of observing any genotypic combination at any locus in terms of the genetic correlation has been derived, the joint likelihood function for all markers can be easily derived. Let N be the number of markers for which genotypic information on a particular pair is unavailable, then the likelihood of observing this pair's genotypes at these loci, assuming independence, is simply the product of the appropriate functions for the pair at each locus.

L (pair genotypes/p_i, R) = $F_i(p_i, R)$, where F_i is the likelihood function at the i^{th} locus with gene frequency vector p_i. The p_i here are treated as known parameters. Dominant marker systems such as ABO and Gm can be handled by summing up the likelihood functions for all possible combinations of pair genotypes compatible with the pair phenotypes.

A computer program to efficiently determine the appropriate function at each locus and calculate the pair likelihood for a given value of R was written, and this was used in conjunction with nonlinear optimization routine QNMDIF [2] to obtain maximum likelihood estimates of R for each pair. In addition, the hypothesis H_0:R=R_0 tested for each pair where R_0 is the theoretical value of R for each type, ie, R_0 = 0.0 for twin-spouse and spouse-spouse correlations and R_0 = 0.5 for twin–twin and sib–sib correlations. The hypothesis test used was the chi-square approximation to the generalized likelihood ratio test (GLR).

Next we computed "lod" scores (as defined by Morton [4] but used here in a nonsequential manner) for each pair:

$$\text{LOD}(R') = \text{Log}_{10}\ \frac{L(\text{pair } R=R')}{L(\text{pair } R=R_0)}$$

for an array of R′ values, R′ = 0, 0.05, 0.1. . . 0.95, 1.0.

https://doi.org/10.1017/S0001566000007765 Published online by Cambridge University Press

These were then summed across all pairs of a particular correlation type in the sample. This was done for each type of correlation in the sample: Twin–Twin (61 pairs), MZ twin–spouse (146 pairs), DZ twin–spouse (122 pairs), MZ spouse–spouse (73 pairs), DZ spouse–spouse (61 pairs), sib–sib (164).

For comparison purposes, theoretical lod score distributions for $R_0=0.0$ and $R_0=0.5$ were generated by looking at all possible pair genotypes at a given marker locus and computing

$$\sum_{i=1}^{N} F_i(p_i R_0) \log_{10} \frac{L(\text{Pair}/R')}{L(\text{Pair } R_0)}$$

for $R' = 0, 0.05, \ldots, 0.95, 1.0$, where N = number of possible pair genotypes.

These are summed across all the markers for each R' to arrive at the theoretical lod score distribution. All heterogeneity tests between lod score distributions were performed using the chi-square test proposed by Morton [5].

RESULTS

The results of this correlational analysis show that the relationships of the theoretical pairs do vary significantly from expectation.

Marriage Pairs

There is a significant correlation between mates in this sample. The Tw–Sp correlation for DZ twins and their mates was estimated to be R=0.086 (Table). This is significantly different from R=0.0 at less than the 0.001 level. MZ twins appeared to marry individuals more similar to themselves and the MZ Tw–Sp correlation was found to be R=0.091. The difference between Tw–Sp correlations for MZ vs DZ twins was not found to be significant, however ($\chi^2 = 0.68$, df = 1). The probability density distributions are shown in Figure 1. The average lod score is plotted against correlation and it is obvious that the distribution for mating pairs is different from the theoretical distribution expected for pairs with a correlation equal to 0. No heterogeneity within the set of mating pairs was found for the MZ or the DZ sample, individually, or for both samples pooled.

It is possible that certain sets of twins may, as a pair, mate with individuals who are more likely to be alike than would randomly selected pairs. The correlation between two presumably random individuals who marry a twin set are called the Sp–Sp or spouse correlation. The distribution of the lod scores for spouse–spouse correlations are shown in Figure 2. While DZ spouse–spouse pairs showed a maximal probability for R=0.060,

TABLE. Estimated Correlations for All Pairs Analyzed in the Present Study

Pair	N	r	lod
Sp–Tw, DZ	119	0.073	3.927
Sp–Tw, MZ	149	0.091	5.025
Sp–Tw, All	268	0.086	7.933
Sp–Sp, DZ	57	0.060	1.254
Sp–Sp, MZ	73	0.111	5.475
Sp–Sp, All	130	0.092	6.472
Random	130	0.006	0.281
Sib–sib	164	0.559	1.752
DZ Twin	61	0.458	1.067

https://doi.org/10.1017/S0001566000007765 Published online by Cambridge University Press

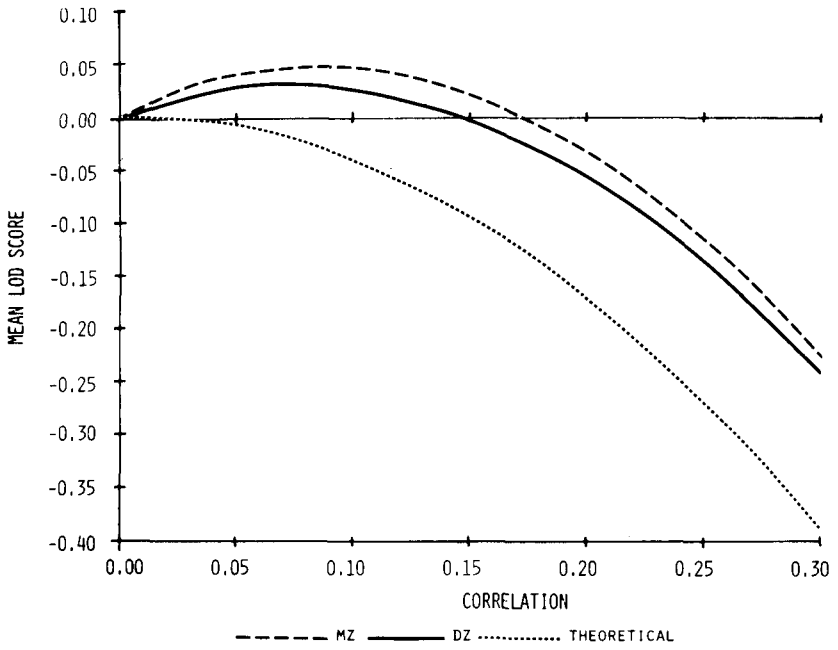

Fig. 1. Distribution of mean lod scores for married pairs.

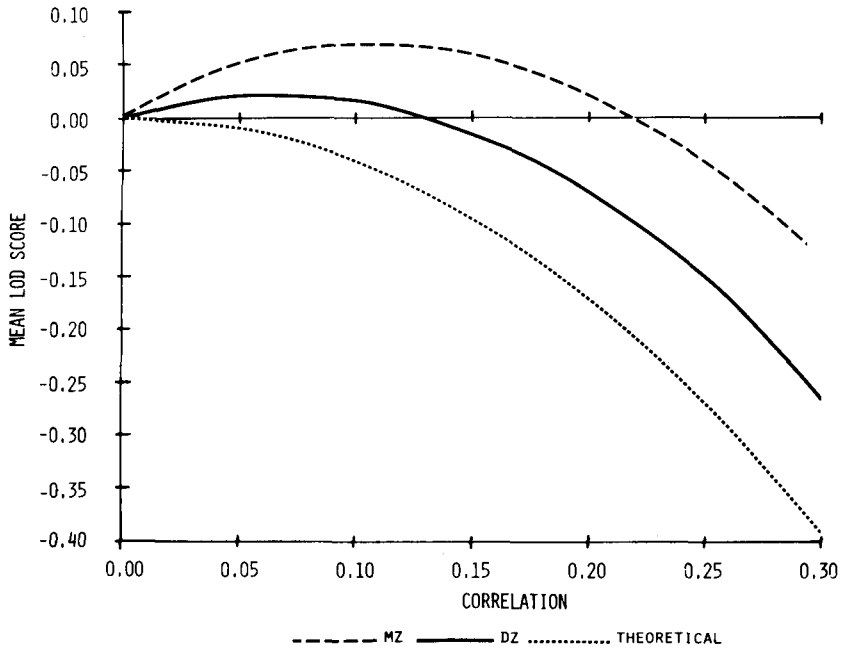

Fig. 2. Distribution of mean lod scores for spouse pairs.

https://doi.org/10.1017/S0001566000007765 Published online by Cambridge University Press

this was not significantly different from R=0. MZ twins, on the other hand, showed a maximal probability for a correlation of R=0.111 ($P < 0.001$). No significant heterogeneity in the probability distributions between DZ and MZ twins was found. The heterogeneity of the lod scores within both the MZ and DZ spouse–spouse sample was examined. It was found that there was no significant heterogeneity within either sample.

In order to test the hypothesis that random pairs should have an average correlation of 0, random pairs of unrelated individuals were generated by a randomizing algorithm. The probability distribution of all for these pairs is shown in Figure 3. The observed probability distribution curve is not significantly different from that expected on the theoretical basis.

Sib and Twin Pairs

Sib-pair correlations in the second generation of our sample were also calculated. A total of 164 pairs were available for analysis. These pairs were not independent because all possible pairs from each of 100 sibships were used. The graphical results are shown in Figure 4 and indicate a maximum probability of R being observed at R=0.559. This however, was not a significant deviation away from the expected 0.5. No significant heterogeneity between sibs in the sample of sib pairs was observed.

Correlation for DZ twins was also calculated. Theoretically, this should parallel that for sib pairs. The probability distribution was found to be maximal for correlation at 0.458 (Fig. 4). No significant heterogeneity between DZ pairs was observed. However, the difference from 0.50 was not significant. Comparison of heterogeneity between the DZ sib pairs and that for ordinary sib paris was found to be significant ($\chi_1^2 = 10.36$, df=1), indicating that the DZ correlation is significantly less than the sib–sib correlation.

CONCLUSIONS

It is not surprising that the genetic correlation for mating pairs estimated by the use of markers was found to be significantly greater than 0. The gene frequencies that were used to calculate the probability distributions were derived from the same sample in which these probability distributions were estimated. The population was assumed to be homogeneous. Any given pair of mates would likely have been drawn from the same small population within Sweden, and the genetic correlation observed probably reflects this fact. It is therefore reasonable to attribute this to nonrandom mating due solely to the phenomenon that individuals tend to marry individuals who live close to them and are genetically more similar. It is surprising, however, that this is observable at the marker gene level and indicates that some of the results relating to assortative mating of continuous variables such as height, intelligence, personality, etc, which are frequently the same magnitude as has been measured here, do not reflect assortative mating in the sense that two individuals seek one another out, but rather represent the effect of a nonrandom distribution of potential marriage partners within the population. In fact, these results indicate that in order to show true assortative mating, the phenotypic correlation must be significantly greater than the correlation between mates calculated on the basis of marker gene results.

We are especially interested in the distribution of correlations of DZ twins. The expectation is that this correlation should be 0.5 and any deviation from this may indicate the presence of either other kinds of biologic twinning or special population phenomena

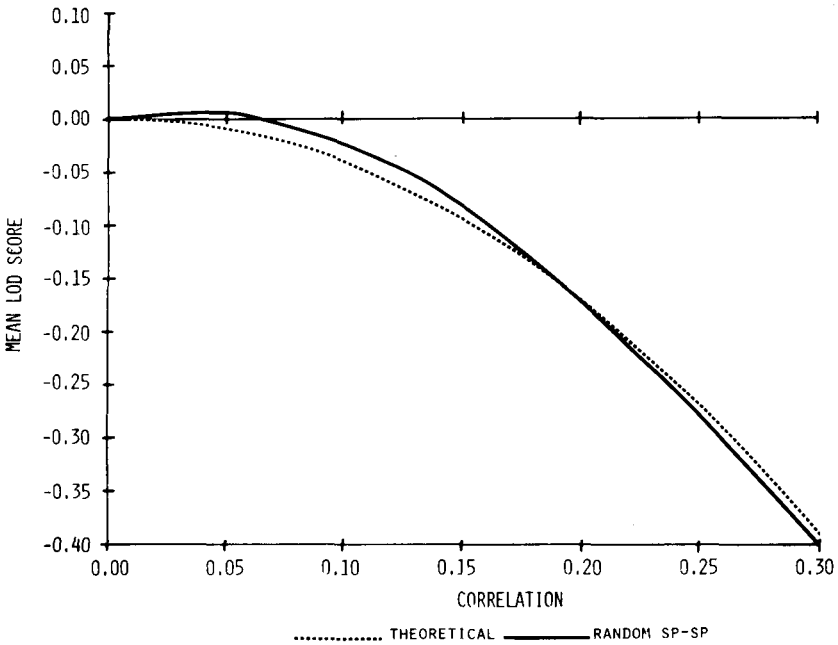

Fig. 3. Distribution of mean lod scores for random pairs in the parental generation.

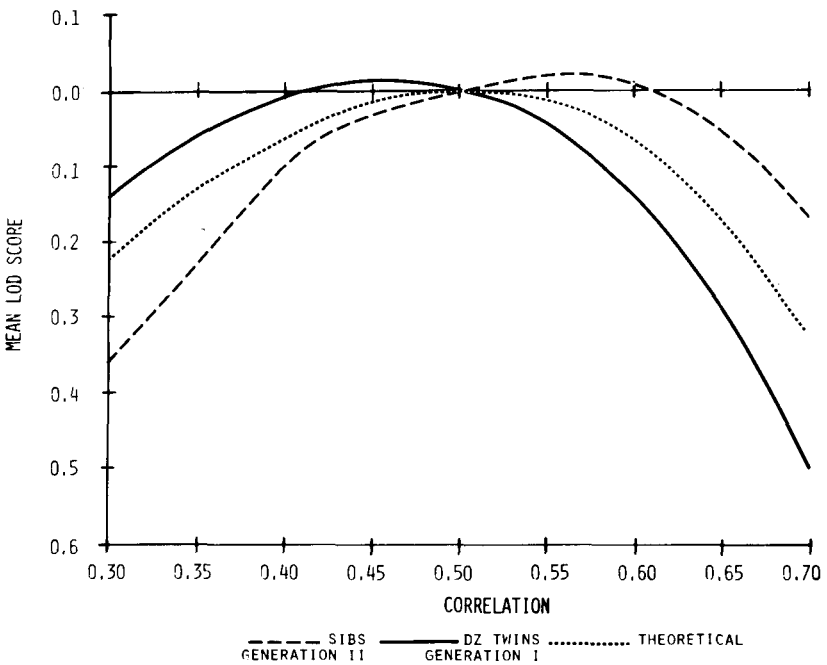

Fig. 4. Distribution of mean lod scores for sib-sib and DZ twin pairs.

https://doi.org/10.1017/S0001566000007765 Published online by Cambridge University Press

such as selection which may tend to make older twins either less or more alike. The DZ twins were all from the older generation, and their ages were between 45 and 60. Instead of comparing the correlation of DZ twins with 0.5, it should be compared with a population of sib pairs. In the present study, these were drawn from the younger generation of 20–35 years of age. A significant difference between these two samples was observed, although neither differs significantly from 0.50. This difference cannot be attributed to differences in mating patterns in the population since the correlation for DZ twins (for the older generation) is significantly lower than that for sib pairs. It would have been expected, with the reasonable assumption of a more highly subdivided population of 40 to 80 years ago (the time during which the parents of the twins in the present study were marrying), that this effect should give rise to a correlation in the DZ sample higher than that of the present generation of sib pairs. Exactly the reverse was observed. It should be pointed out that the correlation for sib pairs was found to be 0.559 which is close to that expected on the basis of the average correlation found between mates in the previous generation of 0.086.

The cause of the significantly lower correlation between DZ twin pairs is at present unknown. One possible explanation is that within the sample there exists a significant number of polar body twins. Such pairs would give correlations less than 0.5. There is, however, as yet no compelling evidence that such twin pairs do exist. Another possible explanation is that for some reason DZ twins who are alike tended to be selected out of the sample either because of ascertainment bias or because of some as yet unknown biological effect.

REFERENCES

1. Crumpacker DW, Cederlof R, Friberg L, Kimberling WJ, Sorenson S, Vandenberg SG, Williams JS, McClearn GE, Grever B, Iyer H, Krier M, Pederson NL, Price RA, Roulette I (1979): A twin methodology for the study of genetic and environmental control of variation in human smoking behavior. Acta Genet Med Gemmellol 28:173–195.
2. Gill, Murray, Pitfield (1972): The implementation of two revised quasi-newton algorithms for unconstrained optimization. National Physical Laboratory Report DNAC 11.
3. Goldgar DE (1981): Statistical methodology for the partitioning of the genetic variance of a quantitative character to specific chromosomes. Unpublished doctoral dissertation, University of Colorado.
4. Morton NE (1955): Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318.
5. Morton NE (1956): The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet 8:80–96.
6. Risch N, Lange K (1979): Application of a recombination model in calculating the variance of sib pair identity. Ann Hum Genet 43:177–186.

Correspondence: W.J. Kimberling, Ph.D., Boys Town Institute, 555 North 30th Street, Omaha, NB, 68131, U.S.A.