Perspectives in Diabetes

Disease Associations

Chance, Artifact, or Susceptibility Genes?

NANCY J. COX AND GRAEME I. BELL

Numerous genes that might contribute to the development of diabetes mellitus and/or its complications have been isolated and characterized. One approach to determining whether these "candidate" genes influence susceptibility to diabetes is to compare the frequency of a DNA marker(s) (restriction-fragment–length polymorphism) for each gene in appropriately matched groups of patients and control subjects. The identification of a DNA-marker association would suggest that genetic variation at this gene may increase or reduce the risk of developing diabetes. However, the absence of an association does not necessarily imply that this gene does not contribute to the development of diabetes. We discuss the genetic rationale of disease association studies and the importance of sample size and disease-marker allele frequencies in these studies.
***Diabetes* 38:947–50, 1989**

The recognition that genetic factors contribute to the development of diabetes has spurred efforts to identify susceptibility genes. The region on human chromosome 6 encoding the human leukocyte antigens (HLAs) was the first to be implicated in diabetes susceptibility, based on a comparison of HLA allele frequencies between patient and control groups. The association of specific HLA markers with insulin-dependent diabetes mellitus (IDDM) susceptibility provided a guide for subsequent family and molecular analyses and illustrates how associations can contribute to the identification of susceptibility genes in a complex disorder (1). Unfortunately, disease-association studies with other candidate genes for IDDM and/or non-insulin-dependent diabetes mellitus (NIDDM), including insulin (2–5), insulin-receptor (6–8), T-lymphocyte–receptor β-chain (9–11), erythrocyte/HepG2 glucose transporter (12,13), and various apolipoprotein (8,14) genes, have not been as informative and, in fact, have often produced conflicting or inconsistent results. Our purpose is not to provide an exhaustive and critical review of association studies on diabetes but rather to consider the biological and statistical issues in the evaluation of such studies.

From the Howard Hughes Medical Institute and the Departments of Medicine and of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois.

Address correspondence and reprint requests to Dr. Nancy J. Cox, Howard Hughes Medical Institute, University of Chicago, 5841 South Maryland Avenue, Box 391, Chicago, IL 60637.

Received for publication 6 February 1989 and accepted in revised form 31 March 1989.

ASSOCIATION-STUDY DESIGN AND RATIONALE

Association studies compare patient and control groups with respect to some marker. Although early disease-association studies used serological, blood group, or enzymatic markers, most recent studies use DNA markers. The methods for typing these DNA markers, called restriction-fragment–length polymorphisms (RFLPs), are straightforward (15). Because many genes involved in carbohydrate and lipid metabolism as well as non-HLA components of the immune system (e.g., T-lymphocyte receptors) have been cloned and characterized with respect to RFLPs, markers are available for various candidate susceptibility genes for both IDDM and NIDDM. Thus, an investigator need only assemble appropriately matched groups of patients and nondiabetic control subjects, isolate DNA, type the individuals, and determine if the marker frequencies differ between the two groups.

The biological rationale for association studies of RFLPs relies heavily on the concept of linkage disequilibrium—the nonrandom association of alleles at different loci. Thus, the underlying supposition of an association study of RFLPs is that any allele frequency differences detected are due to linkage disequilibrium between alleles at the marker locus and alleles at a tightly linked locus that influences susceptibility to disease.

SELECTION OF PATIENT AND CONTROL GROUPS

The most important aspect of an association study is the selection of the patient and control groups. Because RFLP

allele frequencies can vary considerably among ethnic and racial groups, if patient and control groups are not appropriately matched, any observed associations may simply reflect the ethnic, racial, and/or admixture differences between the groups (e.g., 16). Ideally, the patient and control groups should be matched for race, ethnicity, and admixture (if relevant) as well as age, body mass index, and life-style. In practice, however, it is often difficult to match perfectly for each of these variables. As a consequence, some members of the control group are likely to have an unexpressed susceptibility to diabetes, thereby decreasing the ability to detect significant effects.

STATISTICAL ANALYSES

Group differences in a disease-association study may be assessed by comparing the number of alleles of each type in the patient and control groups with a χ^2-test or equivalently by comparing allele frequencies in the two groups with a Z value. This approach is reasonably powerful and requires no assumptions about the mode of transmission of the disease susceptibility locus. Comparing genotype (or pooled genotype) frequencies is not necessary. The genotypic data are not unimportant, however, because they can be used to assess the evidence for mode of transmission of a susceptibility locus (i.e., recessive or dominant) once an association has been established (17).

An additional issue concerns the performance of multiple tests for association. When many loci are tested for association, some will show significant associations by chance. Candidate genes, by definition, have a reasonable prior probability of being involved in disease susceptibility. Correction for the number of candidate genes that might ultimately be tested would be conservative and possibly counterproductive. However, investigators should report the number of candidate genes actually tested so that readers may put results of all studies into a meaningful context.

A further complexity arises when multiple RFLPs have been characterized for a single candidate gene. When six or eight RFLPs are typed for each candidate gene, the number of possible tests becomes large even in studies considering only a few candidate genes. One solution to this difficulty is to consider the information on all RFLPs at a single candidate gene simultaneously by comparing patient and control groups for haplotype frequencies rather than allele frequencies (8). Among the advantages to this approach are that the information on all of the RFLPs is utilized in a biologically relevant manner, whereas the number of tests is reduced to the number of candidate genes being considered.

SAMPLE-SIZE CONSIDERATIONS

Predicting the sample size necessary to detect a statistically significant allelic association requires knowledge of the quantitative contribution of the locus to disease susceptibility, the mode of transmission of the susceptibility locus, and the degree of linkage disequilibrium between the susceptibility and marker loci. Because these variables will always be unknown when an association study is designed, it is virtually impossible to rationally specify sample sizes. However, it is instructive to calculate sample sizes necessary to detect associations with reasonable assumptions for NIDDM disease susceptibility loci (Table 1).

Consider disease susceptibility locus A with alleles A1 and A2 whose frequencies are q_{A1} and q_{A2} (where $q_{A1} + q_{A2} = 1$), and let A1 be the allele that increases susceptibility to diabetes. The penetrances for the three genotypes are denoted f_{A1A1}, f_{A1A2}, and f_{A2A2} and are defined as the probability of being affected given that genotype (A1A1, A1A2, or A2A2, respectively). The cumulative incidence of the disease in a population can be denoted as K_p and may be calculated from the penetrances and allele frequencies as

$$K_p = q_{A1}^2 f_{A1A1} + 2q_{A1}q_{A2}f_{A1A2} + q_{A2}^2 f_{A2A2}$$

In all the examples we considered, we fixed K_p at .05 and have allowed for the disease susceptibility allele to be dominant but incompletely penetrant (i.e., $f_{A1A1} = f_{A1A2} < 1$). In addition, the susceptibility locus accounts for only a proportion of the affected individuals, denoted Pr_{Aff}, and is calculated as

$$\frac{q_{A1}^2 f_{A1A1} + 2q_{A1}q_{A2}f_{A1A2}}{K_p}$$

TABLE 1
Effects of contribution of susceptibility locus and level of linkage disequilibrium on sample size required to detect association

	Example						
	1	2	3	4	5	6	7
f_{A1A1}	.95	.95	.95	.95	.95	.15	.95
f_{A1A2}	.95	.95	.95	.95	.95	.15	.95
f_{A2A2}	.035	.035	.035	.035	.04	.04	.01
q_{A1}	.008	.008	.008	.008	.005	.05	.021
Pr_{Aff}	.30	.30	.30	.30	.20	.30	.80
q_{B1}	.10	.10	.20	.10	.10	.10	.10
D'	+.80	−.80	+.80	+.40	+.80	+.80	−.80
$q_{B1}(D)$	.205	.088	.294	.152	.167	.174	.068
$n(.5)$	44	2415	80	154	97	82	291
$n(.8)$	89	4933	164	314	197	167	595

f_{A1A1}, f_{A1A2}, f_{A2A2}, Penetrances for genotypes at susceptibility locus; q_{A1}, susceptibility allele A1 frequency; Pr_{Aff}, proportion of affected individuals with disease due to susceptibility locus A; q_{B1}, marker allele B1 frequency for general population; D', linkage disequilibrium between susceptibility and marker loci; $q_{B1}(D)$, marker allele B1 frequency expected in disease population; $n(.5)$, sample size (number of individuals) required to detect difference between q_{B1} and $q_{B1}(D)$ with probability of .5, assuming significance level of .05; $n(.8)$, sample size required to detect difference between q_{B1} and $q_{B1}(D)$ with probability of .8, assuming significance level of .05.

Thus, the proportion of phenocopies, i.e., individuals who have disease due to other loci or environmental factors, is $1 - Pr_{Aff}$, or $q_{A2}^2 f_{A2A2}/K_p$. Locus B is a marker locus (e.g., an RFLP from a noncoding region detected with a probe for a candidate gene, locus A) in linkage disequilibrium with locus A; locus B has alleles B1 and B2 with frequencies q_{B1} and q_{B2} in the general population. Linkage disequilibrium may be measured with the value D', which provides an indication of the linkage disequilibrium relative to its theoretical maximum for a given pair of loci (18). Table 2 illustrates the calculation of D' and provides an example of the expected frequencies of gametes with A1B1, A1B2, A2B1, and A2B2 under various levels of disequilibrium between the A and B loci. Once the penetrances and allele frequencies for susceptibility locus A, the allele frequencies for the marker locus B in the general population, and the degree of linkage disequilibrium between the two loci are specified, it is possible to calculate the marker allele frequencies expected for a disease population, which are denoted $q_{B1}(D)$ and $q_{B2}(D)$, and the sample sizes (n) necessary to detect the difference between allele frequencies for marker locus B in the general population and in the disease population. Seven examples are presented in Table 1.

Example 1. We consider a dominant susceptibility locus with relatively high penetrance but low susceptibility allele frequency, such that only 30% of affected individuals have disease due to locus A. This gene represents a major susceptibility factor but only in a few families. The marker locus allele B1 is found with a frequency of .1 in the general population and is associated with the disease susceptibility allele A1 ($D' = +.8$). Therefore, the B1 allele frequency expected for a population of individuals with disease is .205. Assuming a significance level of .05, the sample size required to detect this association, i.e., the difference between the B1 allele frequency in the general population and the B1 allele frequency in the disease population, with a probability of .5 is 44 and with a probability of .8 is 89. That is, if the true frequencies of B1 in the general and disease populations were .1 and .205, respectively, we would detect a significant difference in the B1 allele frequency between the two groups 50% of the time in a random sample of 44 patients and 44 control subjects and 80% of the time if we collected data on a random sample of 89 patients and 89 control subjects.

TABLE 2
Haplotype frequencies expected for various levels of linkage disequilibrium

	Haplotype frequencies expected for:		
	$D' = 0$	$D' = +.8$	$D' = -.8$
A1B1	.005	.041	.001
A1B2	.045	.009	.049
A2B1	.095	.059	.099
A2B2	.855	.891	.851

We assumed 2 loci, A and B, each with 2 alleles, A1 and A2 and B1 and B2. Allele frequencies were assumed to be $q_{A1} = .05$, $q_{A2} = .95$, $q_{B1} = .1$, and $q_{B2} = .9$. D' is calculated as $D' = D/D_{max}$, where $D = h_{A1B1} - q_{A1}q_{B1}$ (where h_{A1B1} is observed frequency of haplotypes with A1 and B1 alleles) and $D_{max} = \min(q_{A1}q_{B1}, q_{A2}q_{B2})$ for $D < 0$ or $\min(q_{A1}q_{B2}, q_{A2}q_{B1})$ for $D > 0$.

Example 2. All parameters are the same as in example 1 except the B1 allele is associated with allele A2 at the susceptibility locus; therefore, $D' = -.8$ instead of $+.8$ as in example 1. Because the common allele at one locus is associated with the rare allele at a second locus, there is much less deviation from the haplotype frequencies expected under no disequilibrium than when the rare alleles at each locus are associated; therefore, the sample size required to detect the difference is large.

Example 3. The parameters for the disease susceptibility locus are the same as in examples 1 and 2, but the marker allele frequency in the general population, q_{B1}, is increased to .2, and the B1 and A1 alleles are associated ($D' = +.8$). Under these assumptions, the frequency of the B1 allele expected for a disease population is .294, and the sample sizes required to detect this difference are almost twice those required for example 1. Thus, the more similar the frequencies of the associated alleles, the easier it is to detect the linkage disequilibrium between the loci, and therefore the easier it is to detect the association. Even though the magnitude and the sign of linkage disequilibrium between the A and B loci were the same for examples 1 and 3, in example 1, an allele frequency .008 was associated with an allele with frequency .1, whereas in example 3, an allele with frequency .008 was associated with an allele with frequency .2, accounting for the difference in sample sizes required to detect the associations for these examples.

Example 4. The parameters for the susceptibility locus are the same, and we return the B1 allele frequency in the general population to .1, but we reduce the linkage disequilibrium between the A and B loci to $D' = +.4$. Under these conditions, the B1 allele frequency expected for the disease population increases to only .152, compared with the $q_{B1}(D)$ of .205 from example 1 where $D' = +.8$. Therefore, we require more than triple the sample size of example 1 to detect the association produced in example 4.

Example 5. We consider the effect of reducing the proportion of affected individuals who have disease due to locus A. Although we retain the high penetrances for the susceptible homozygotes and heterozygotes, we reduce the proportion of affected individuals with disease due to locus A from 30% (examples 1–4) to 20% by reducing q_{A1} from .008 to .005 and slightly increasing the penetrance in normal homozygotes. Keeping the $q_{B1} = .1$ and $D' = +.8$, we find the expected frequency of B1 in the disease population is .167, and the sample sizes required to detect the association are more than twice those required for example 1.

Example 6. We demonstrate that not only the magnitude of the contribution from the susceptibility locus is important in the ability to detect an association but also how the contribution is made. The penetrances for the susceptible homozygote and heterozygote are reduced in example 6, but the susceptibility allele frequency is increased, so that the proportion of affected individuals with disease due to susceptibility from locus A is still 30%. The susceptibility locus in this example could represent one of a few common minor genes that interact to produce disease. Again, keeping $q_{B1} = .1$ and $D' = +.8$, the expected frequency of B1 in the disease population is .174. Comparing the sample sizes required to detect the association in example 6 with those required to detect the association in example 1, it is clear

that the reduced penetrances decrease the ability to detect the association, even though the locus accounts for the same proportion of affected individuals in both examples.

Example 7. The susceptibility locus is the major factor accounting for disease in 80% of affected individuals and is therefore the major gene for disease susceptibility. However, because the susceptibility allele is associated with the common allele at the marker locus, relatively large sample sizes are needed to detect the association. Example 6 demonstrates how it is possible to detect significant associations with modest sample sizes and yet fail to find evidence for linkage. Conversely, example 7 demonstrates how it is possible for a major susceptibility locus, detectable through linkage studies, to show no evidence of association with tightly linked markers.

INTERPRETING RESULTS OF ASSOCIATION STUDIES

Critical evaluation of the results of an association study with RFLP markers requires consideration of many factors. Possible interpretations include chance, artifact, or that the marker examined is in linkage disequilibrium with variation that affects disease susceptibility. It is important to adequately describe the patient and control populations and to indicate how many tests for association were performed. Finally, it is important to emphasize that negative results (failure to detect an association) are not conclusive, because there may have been insufficient power to detect the contribution of that locus to disease susceptibility. Positive results are consistent with the possibility that a gene(s) in the marker region affects disease susceptibility. However, it is usually not possible from the association study to determine whether the susceptibility locus is a major or minor contributor to disease; linkage studies in families may aid in making that determination.

Are association studies worthwhile? We believe so, at least when properly designed and cautiously interpreted. Association studies should be considered exploratory and preliminary, with results requiring confirmation. It is possible that association studies will identify subgroups within NIDDM, just as the HLA association distinguishes IDDM and NIDDM. Such clarification of heterogeneity would be invaluable to family and linkage studies of NIDDM. Moreover, the possibility that NIDDM is largely multifactorial and/or polygenic, with rare major genes contributing most of the susceptibility in only a few families, cannot be rejected. If this is the case, association studies may be the only way of identifying the "polygenes" that contribute susceptibility to NIDDM.

ACKNOWLEDGMENTS

We thank Julie Dicig for assistance in the preparation of this manuscript and Drs. Carole Ober, Tim Shapiro, and Paul Epstein for helpful comments.

REFERENCES

1. Tiwari JL, Terasaki PI (Eds.): *HLA and Disease Associations*. New York, Springer, 1985
2. Owerbach D, Nerup J: Restriction fragment length polymorphism of the insulin gene in diabetes mellitus. *Diabetes* 31:275–77, 1982
3. Bell GI, Horita S, Karam JH: A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33:176–83, 1984
4. Hitman GA, Tarn AC, Winter RM, Drummond V, Williams LG, Jowett NI, Bottazzo GF, Galton DJ: Type 1 (insulin-dependent) diabetes and a highly variable locus close to the insulin gene on chromosome 11. *Diabetologia* 28:218–22, 1985
5. Cox NJ, Baker L, Spielman RS: Insulin-gene sharing in sib pairs with insulin-dependent diabetes mellitus: no evidence for linkage. *Am J Hum Genet* 42:167–72, 1988
6. Elbein SC, Corsetti L, Ullrich A, Permutt MA: Multiple restriction fragment length polymorphisms at the insulin receptor locus: a highly informative marker for linkage analysis. *Proc Natl Acad Sci USA* 83:5223–27, 1986
7. McClain DA, Henry RR, Ullrich A, Olefsky JM: Restriction-fragment–length polymorphism in insulin-receptor gene and insulin resistance in NIDDM. *Diabetes* 37:1071–75, 1988
8. Xiang K-S, Cox NJ, Sanz N, Huang P, Karam JH, Bell GI: Insulin-receptor and apolipoprotein genes contribute to development of NIDDM in Chinese Americans. *Diabetes* 38:17–23, 1989
9. Hoover ML, Angelini G, Ball E, Stastny P, Marks J, Rosenstock J, Raskin P, Ferrara GB, Tosi R, Capra JD: HLA-DQ and T cell receptor genes in insulin dependent diabetes mellitus. *Cold Spring Harbor Symp Quant Biol* 51:803–809, 1986
10. Millward BA, Welsh KI, Leslie RDG, Pyke DA, Demaine AG: T cell receptor beta chain polymorphisms are associated with insulin-dependent diabetes. *Clin Exp Immunol* 70:152–57, 1987
11. Spielman RS, Cox NJ, Syvester DR, Concannon P: Segregation of T-cell receptor beta chain RFLPs in sibs with insulin-dependent diabetes mellitus (IDDM) (Abstract). *Am J Hum Genet* 41:A186, 1987
12. Li SR, Baroni MG, Oelbaum RS, Stock J, Galton DJ: Association of genetic variant of the glucose transporter with non-insulin-dependent diabetes mellitus. *Lancet* 2:368–70, 1988
13. Cox NJ, Xiang KS, Bell GI, Karam JH: Glucose transporter gene and non-insulin-dependent diabetes. *Lancet* 2:793–94, 1988
14. Buraczynska M, Hanzlik J, Grzywa M: Apolipoprotein A-I gene polymorphism and susceptibility of non-insulin-dependent diabetes mellitus. *Am J Hum Genet* 37:1129–37, 1985
15. Caskey CT: Disease diagnosis by recombinant DNA methods. *Science* 236:1223–29, 1987
16. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: $Gm^{3;5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 43:520–26, 1988
17. Thomson G: Investigation of the mode of inheritance of the HLA associated disease by the method of antigen genotype frequencies among diseased individuals. *Tissue Antigens* 21:81–104, 1983
18. Lewontin RC: The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* 49:49–67, 1964