Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: Studies of the HLA class II loci

(restriction fragment length polymorphism/diabetes)

NORMAN ARNHEIM, CAROLYN STRANGE, AND HENRY ERLICH

Human Genetics Department, Cetus Corporation, 1400 Fifty-Third Street, Emeryville, CA 94608

Communicated by L. L. Cavalli-Sforza, June 6, 1985

ABSTRACT A rapid method has been developed and used to search for restriction fragment length polymorphisms (RFLPs) that are in linkage disequilibrium with disease-associated loci. By using genomic blot-hybridization analysis with DQ β-chain and DR β-chain cDNA probes, we examined DNA polymorphisms within the HLA class II loci associated with susceptibility to insulin-dependent diabetes mellitus (IDDM). To facilitate the search for informative RFLPs, we compared pooled DNA samples from IDDM patients with pooled DNA samples from randomly selected control individuals, instead of using the conventional approach of examining DNA samples from individuals in two groups. (The conditions under which this approach is useful are treated theoretically in the Appendix.) Several specific polymorphic restriction fragments associated with IDDM were revealed by using this economical and rapid approach. The restriction enzymes and probes identified as informative in this screening were then used to analyze HLA-DR-typed IDDM families, homozygous typing cells, and unrelated individuals to determine the association of the specific restriction fragments with the HLA-DR serological type and the frequency in control and IDDM populations. Some individual polymorphic fragments for which the IDDM population was enriched correlated strongly with HLA-DR3, whereas others correlated strongly with HLA-DR4. Some fragments (e.g., a 10-kilobase Taq I fragment detected with the DRβ probe) that were more prevalent in the IDDM population subdivided the serologically defined HLA-DR type and may be informative markers for IDDM susceptibility.

DNA polymorphisms tightly linked to loci at which mutations cause disease or disease susceptibility can serve as useful genetic markers to elucidate modes of inheritance and as a tool for genetic diagnosis. If a particular polymorphic variant is in linkage disequilibrium with the disease-related locus, then the DNA polymorphism can be informative beyond its value as a marker in linkage analysis, as was the association between sickle-cell anemia and a 13-kilobase (kb) Hpa I restriction fragment of the β-globin gene (1).

Genetic susceptibility to a number of diseases shows linkage as well as association in population studies with specific serologically defined variants of the HLA class I and class II loci (reviewed in ref. 2). These loci encode cell surface glycoproteins that mediate a variety of immune functions. The association of many diseases that have an autoimmune component, such as insulin-dependent diabetes mellitus (IDDM), with serologically defined class II variants (HLA-DR types) suggests the existence of disease-susceptibility-related genes in linkage disequilibrium with specific serologic markers. It is not known whether the class II gene products are themselves involved in disease causation and, if so, whether the variants of the polymorphic sequences that form the basis of immunologic typing are the same ones that confer disease susceptibility.

The genetic basis of IDDM susceptibility is complex; 60–80% of the genetic component is thought to be HLA-linked (3), and the penetrance, as revealed by monozygotic twin studies, is about 50% (4). IDDM has been associated in population studies with the serologically defined DR3 and/or DR4 specificities; about 90% of IDDM patients are either DR3 and/or DR4, compared to about 55% of control individuals (5). The relatively low proportion of DR3 and DR4 individuals who develop IDDM may reflect the incomplete penetrance as well as the apparent genetic heterogeneity (6–8) within the serologically defined DR types, such as DR3 and DR4.

The HLA class II gene products, as defined by serologic and cellular typing techniques, consist of the HLA-DR, HLA-DQ, and HLA-DP antigens, heterodimers comprised of an α (32–34 kDa) and a β (28–29 kDa) chain (reviewed in ref. 9). The genetic organization of the class II region, as revealed by molecular analysis with cDNA and genomic clones, is considerably more complex, however. The number of loci encoding the class II α and β chains may be as great as 13 (9–13). In addition, serologically indistinguishable DR types show allelic variation at the DNA level (refs. 8 and 14–16; unpublished results). HLA DNA polymorphisms, detected as polymorphic restriction fragments, promise to be a rich source of well-defined genetic markers for the analysis of HLA organization and the relationship between class II loci and disease susceptibility.

Accordingly, considerable effort has been made to detect DNA markers in linkage disequilibrium with HLA-linked diseases (8, 15, 16) in hope of increasing the utility of the HLA markers in predicting susceptibility to diseases. The standard procedure for identifying informative DNA polymorphisms utilizes a cloned HLA sequence as a hybridization probe and compares the restriction fragment patterns obtained with a variety of restriction enzymes in Southern blots (17) of DNA samples from diseased and normal individuals. This analysis represents a search for (i) enzymes that reveal polymorphisms and (ii) those specific fragments whose frequency is increased in the patient population. This screening procedure is long, laborious, and consumes large amounts of DNA. We have devised a short-cut method for identifying polymorphic restriction fragments in linkage disequilibrium with disease-related loci that also requires much less genomic DNA. The method is based on the use of pooled preparations of DNA derived from individual blood samples, thereby creating a "control" and a "disease" DNA pool. If restriction digests of these pools are analyzed by Southern blotting in adjacent lanes, a polymorphic fragment that is in linkage disequi-

Abbreviations: IDDM, insulin-dependent diabetes mellitus; RFLP, restriction fragment length polymorphism; kb, kilobase(s).
librium with the putative disease-related locus will be revealed as a band that is more intense in the disease pool than in the control pool. Thus, this search for informative restriction endonucleases and probes identifies only those fragments that are both polymorphic and increased in frequency among the patient population. Using this approach, we have identified several restriction fragment length polymorphisms (RFLPs), located in or around the DQ β-chain (DQβ) or DR β-chain (DRβ) class II loci, that are in linkage disequilibrium with the gene(s) conferring susceptibility to IDDM. We also discuss the general utility of this approach for the discovery of RFLPs in linkage disequilibrium with other genetic loci.

**MATERIALS AND METHODS**

Southern Transfer and Hybridization. High molecular weight DNA was prepared from blood samples or cell lines from HLA-DR-typed individuals by use of published procedures (18). DNA pools were constructed by mixing equal amounts of DNA from either 31 unrelated control individuals or 21 unrelated diabetic individuals. In each experiment, 10 μg of DNA from the control or diabetic pool was digested with various restriction enzymes under conditions suggested by the supplier (New England Biolabs). Special care was taken to ensure that all digestions were complete. The digested DNAs were ethanol-precipitated and resuspended in 10 mM Tris Cl, pH 7.9/1 mM EDTA before electrophoresis (35 V, overnight) in 17-cm-long agarose gels. The agarose concentration was either 1.3% (experiments with Msp I, Taq I, and Rsa I) or 0.7% (for the remaining enzymes). Southern transfer to Genetran filters, nick-translation, hybridization, and washing procedures were carried out as described (18). Nick-translated plasmids containing DQβ, DRβ, or DRα cDNA inserts (18, 19) had specific activities between 10⁸ and 10⁹ cpm/μg. Autoradiograms were exposed 7–10 days at −70°C, using Kodak XAR-5 film with intensification screens.

**RESULTS**

Our approach toward the discovery of informative HLA polymorphisms was to compare the fragment-intensity patterns of pooled DNA samples from normal and IDDM individuals subjected to restriction enzyme digestion, Southern blotting, and hybridization with various HLA class II cDNA probes. To test whether the pool method could detect a specific polymorphic fragment whose frequency was already known to be elevated in the IDDM patient population (8, 14), the analysis of DNA pools was carried out initially using Bgl II and the DRα probe. Polymorphism in two Bgl II restriction sites located at the 3' end of the HLA-DRα locus gives rise to three allelic variants: a 3.8-, a 4.2-, and a 4.5-kb Bgl II fragment (8). In both the control and the IDDM pools, all three fragments are observed; however, it is clear that the diabetic pool is significantly enriched in the 4.2-kb fragment (Fig. 1a), as would be expected from previous analyses of individual DNA samples from IDDM patients and controls (refs. 8 and 14 and unpublished data).

The pooled-DNA method, using DRβ and DQβ cDNA probes, then was used to search for polymorphisms that are linked to the DRβ and DQβ loci and whose frequency is increased in the IDDM population. The results obtained with some of the enzymes tested are shown in Fig. 1b (DQβ cDNA probe) and Fig. 1c (DRβ cDNA probe). The presence of multiple bands in each lane is explained by several factors, in addition to genetic polymorphism at a single locus as was the case for the DRα study (Fig. 1a). Multiple cross-hybridizing loci encoding either DR β chains (3 loci) or DQ β chains (2 loci) have been detected (12, 13). In addition, since the hybridization probes are cDNA clones, multiple hybridizing fragments would be expected even from a single locus, if its introns contain recognition sequences for the chosen restriction enzyme. Pooled DNA samples digested with 14 different restriction enzymes (BamHI, EcoRI, HindIII, Bgl II, Kpn I, EcoRV, Pst I, Stu I, Sac I, Xba I, Pvu II, Msp I, Taq I, and Rsa I) were hybridized with DQβ and DRβ cDNA probes. The autoradiographic intensities of most of the restriction

---

**Fig. 1.** Southern blot analysis of pooled genomic DNA from 31 control (lanes C) and from 21 diabetic (lanes D) individuals, using HLA class II cDNA probes. (a) The Bgl II, DRα polymorphism. The sizes of the three allelic restriction fragment variants are shown. (b) DQβ restriction fragment patterns with seven restriction enzymes. (c) DRβ restriction fragment patterns with three restriction enzymes. In b and c, the sizes of the fragments can be estimated from the markers.
fragments were not significantly different in the control and IDDM lanes. For any individual band, equivalent intensities reflect the presence of either an invariant fragment in the population or a polymorphic fragment in linkage equilibrium with IDDM. Such a polymorphic fragment would not be an informative genetic marker for IDDM susceptibility, beyond its use in linkage analysis. Some enzymes generated bands that were more intense in the control than in the IDDM pool (e.g., Xba I, with the DQB probe). The analysis with the DQB probe of BamHI-digested DNA pools (data not shown) revealed a 3.7-kb fragment present in the control but not the IDDM lane, consistent with a previously reported observation using individual samples (15). These polymorphic fragments whose frequencies are elevated in the control pool represent potentially useful markers for IDDM susceptibility; however, in the absence of a corresponding allelic-fragment variant that is increased in the IDDM pool, it is difficult to determine the familial segregation pattern and population distribution of the polymorphic restriction site. Intensity differences between control and diabetic pools for the EcoRI and Pst I fragments previously reported to be increased among IDDM patients (15) were not observed in our samples.

The pool analysis did, however, reveal several other restriction fragments whose intensities were increased in the IDDM pools (e.g., 1.5-kb and 2.7-kb Rsa I fragments, detected with the DQB probe). Since these polymorphic fragments represent potential new markers for IDDM susceptibility, their distributions in populations of randomly selected control individuals, IDDM patients, and HLA-DR-typed individuals, including donors of homozygous typing cells (HTCs), were examined. The genomic-blot patterns obtained with the DQB probe and Rsa I digests from a panel of homozygous-typing cells (Fig. 2a) and HLA-DR-typed control individuals (Fig. 2b and data not shown) indicated that all HLA-DR3 individuals (38/38) contain a 2.7-kb Rsa I fragment that hybridizes with the DQB probe; the fragment was not observed in DNA from 52 non-DR3 individuals. [In Fig. 2, the faint band (slightly <2.7 kb) seen in the DR7-typed homozygous typing cell of donor LBF probably represents a different genomic fragment of comparable length.] A 1.5-kb Rsa I fragment that hybridized to the DQB probe was present in most DR4 individuals (13/17 control and 27/30 IDDM); the other DR4 individuals all contained an allelic 1.8-kb fragment. In the normal samples examined, no non-DR4 individuals (0/16) contained the 1.5-kb fragment; however, we have observed the 1.5-kb fragment in DNA derived from four DR7 IDDM patients. The segregation of polymorphic DQB-related Rsa I fragments in 11 IDDM families showed that the 2.7-kb Rsa I fragment segregated with the DR3 haplotype, whereas the 1.5-kb (or 1.8-kb, in some of the families) fragment segregated with the DR4 haplotype. The results obtained with one representative family are shown in Fig. 3.

Since the pool analysis with the DRB probe and Taq I had also revealed potentially informative polymorphisms, an analysis of HLA-DR-typed individuals and families was also carried out. The 7.3-kb DRB-related Taq I fragment was found in 27 out of 27 DR3 individuals and in only 7 out of 26 non-DR3 types, whereas the 2.6-kb Taq I fragment is found in 28 out of 28 DR4 individuals and in only 8 out of 27 non-DR4 individuals. The 10-kb Taq I fragment subdivides DR3 and DRw6 haplotypes, two serologically related types, and may represent a new, informative marker for IDDM susceptibility deserving of further testing, since it is present in most IDDM DR3 haplotypes (9/10) but only about half (6/11) of control DR3 haplotypes.

DISCUSSION

Our analysis of DNA pools with the DQB and DRB probes revealed several new polymorphic fragments that were increased in the IDDM population. These fragments are candidates for genetic markers in linkage disequilibrium with the putative IDDM susceptibility loci. New DNA markers could prove more informative than the serological DR3 and DR4 markers if (i) they subdivide existing serological types or (ii) they identify IDDM-susceptible individuals who are neither DR3 nor DR4. The distribution and family segregation of some of the DQB and DRB polymorphic restriction fragments was examined in samples from individuals and was found to be correlated with either DR3 or DR4, consistent with the idea that the susceptibility loci associated with DR3 and DR4 may be different (2). These markers may prove more informative than the serological ones.

Thus, the DQB-related 1.5-kb Rsa I fragment identifies a number of non-DR4 IDDM individuals as well as 90% of all IDDM DR4 individuals. In addition, the 10-kb Taq I fragment

---

**Fig. 2.** Genomic blot analysis of Rsa I-digested DNA from a panel of HLA-DR-typed individuals, using the DQB probe. (a) Homozygous typing cells (HTCs). Positions of DR3-associated 2.7-kb and DR4-associated 1.5-kb fragments are shown. Each lane is labeled with the serological DR type and the HTC designation. One HTC (MWF) is nonconsanguineous and appears to be heterozygous for many class II loci. (b) Eight unrelated, HLA-DR-typed individuals. The HLA-DR type is given above each lane.
subdivides the DR3 serological type and may represent a new and useful marker for IDDM susceptibility. Because it is 100% correlated with DR3 specificities, the DQβ-related 2.7-kb Rsal fragment cannot provide a better assessment of relative risk (2) for IDDM than can the DR3 marker. This fragment, however, could function as a specific marker for the DR3 type, with which it appears to be correlated absolutely. The Rsal I and Taq I sites that define these DQβ and DRβ polymorphic fragments have been localized with genomic clones from IDDM individuals.

Studies that incorporate an additional control pool matched with the IDDM pool for DR serological specificities would be expected to reveal fragments whose distribution is different from that of the DR3 and DR4 types and should provide markers unique to chromosome 6 of IDDM-susceptible individuals regardless of DR specificity.

The discovery of RFLPs in linkage disequilibrium with a locus involved in disease causation or susceptibility can be laborious and time-consuming. Examination of 20 different restriction enzyme digests from each of 30 normal and 30 diseased individuals with 10 DNA probes, assuming each Southern blot were used only one time, would require 1200 blots (with 10 lanes each). In addition, if 5 µg from each individual DNA sample were used each time, a total of 1 mg of DNA from every individual to be screened would be needed. In comparison, analysis of the same number of individuals, with the same numbers of probes and enzymes, for informative polymorphisms by the pool method would require only 1/N times as much effort and DNA, where N is equal to the pool size.

In the Appendix to this paper, the conditions under which this approach is useful are treated theoretically. It is shown that, at various reasonable levels of disequilibrium (normalized disequilibrium >0.06) between a marker locus and the disease-related locus, the hybridization intensity of the marker fragment in the diseased pool will be at least 3 times greater than that in the control pool, provided that the allelic frequency of the polymorphic marker locus in the population is within an appropriate range (0.01–0.3). In general, the lower the marker-allele frequency in controls, the easier it is to detect differences between the control and disease pools. These considerations make it likely that the pool approach can be applied to the search for RFLPs in linkage disequilibrium with other disease-related loci involving limited molecular heterogeneity, once probes closely linked to those loci are available.

**APPENDIX**

**Glenys Thomson**

Department of Genetics, University of California, Berkeley, CA

A two-locus, two-allele model is considered. (The results may be extended naturally to multiallelic models.) The first “locus”, M, denotes the restriction fragment length polymorphism (RFLP), and thus is the marker locus. The two alleles M and m denote, respectively, the presence or absence of a particular RFLP. The second locus, D, with alleles D and d, is the disease-susceptibility-related locus. There are four gametic or chromosome types possible, namely MD, Md, mD, and md. The frequencies of these gametic types can be expressed in the well-known formulation in terms of allele frequencies and linkage disequilibrium, as illustrated in Eqs. 1. In our case, a second notation, in terms of allele frequencies and the ratio of marker presence to absence (k/(1 - k)) in gametes containing the disease-related allele D, is the appropriate one to use (20, 21). This notation is also given in Eqs. 1:

\[
\begin{align*}
 f(MD) &= p_M p_D + \Delta = k p_D \\
 f(Md) &= p_M q_D - \Delta = p_M - k p_D \\
 f(mD) &= q_M p_D + \Delta = (1 - k) p_D \\
 f(md) &= q_M q_D + \Delta = q_M - (1 - k) p_D,
\end{align*}
\]

where \( p_M \) and \( p_D \) are the frequencies of the alleles M and D, respectively; \( p_M + q_M = 1 = p_D + q_D \); and \( \Delta = f(MD) - f(Md)f(md) \) is the coefficient of linkage disequilibrium. The two notations are completely interchangeable, with

\[ k = p_M + \Delta / p_D. \]

The most general model of disease susceptibility allows all three genotypes DD, Dd, and dd to be disease-susceptible, with penetrance values for the three genotypes denoted \( f_3, f_1, \) and \( f_2 \) (22). We restrict our consideration to intermediate models with \( f_0 = 0 \). For strict recessive models, \( f_1 = f_0 = 0; \) for additive models, \( f_2 = 2 f_1; \) for dominant models \( f_3 = f_1. \) (Since the results for dominant and additive models are very similar when the disease-predisposing allele frequency is not too large, we restrict our attention to the additive model, as the calculations are simpler in this case.)

By using the calculations of Thomson (21), it can be shown that the allele frequency of the marker restriction fragment in patients is given by

\[
\frac{[\alpha p_D + k + p_M]}{[\alpha p_D + 2 p_D]},
\]

where \( \alpha = (f_2 - 2 f_1)/f_1. \) This function is a monotonically increasing function of \( \alpha; \) thus we consider only the two extreme cases of \( \alpha = 0 \) (additive) and \( \alpha = \infty \) (recessive). The marker-allele frequency in patients in the additive case is \( (k + p_M)/2 \), whereas for the recessive case it is \( k. \)

Using Eq. 2 and function 3, we see that the ratio, \( R, \) of the marker restriction fragment-allele frequency in patients versus controls is given by

\[
R(\text{recessive}) = 1 + \Delta / p_M p_D,
\]

or

\[
R(\text{additive}) = 1 + \Delta / 2 p_M p_D.
\]
Table 1. Ratio ($R$) of the frequency of the marker restriction fragment in patients to that in controls, for the recessive model

<table>
<thead>
<tr>
<th>$P_M$</th>
<th>$\Delta'$</th>
<th>1.0</th>
<th>0.5</th>
<th>0.25</th>
<th>0.125</th>
<th>0.0625</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>100</td>
<td>50.5</td>
<td>25.75</td>
<td>13.375</td>
<td>7.1875</td>
<td>3.625</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>10.5</td>
<td>5.75</td>
<td>3.75</td>
<td>2.1875</td>
<td>1.0625</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>5.5</td>
<td>3.25</td>
<td>2.1250</td>
<td>1.5625</td>
<td>0.8333</td>
</tr>
<tr>
<td>0.2</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>1.25</td>
<td>-</td>
</tr>
<tr>
<td>0.3</td>
<td>3.333</td>
<td>2.167</td>
<td>1.5833</td>
<td>1.2917</td>
<td>1.1458</td>
<td>-</td>
</tr>
<tr>
<td>0.333</td>
<td>3</td>
<td>2.0</td>
<td>1.5</td>
<td>1.25</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.4</td>
<td>2.5</td>
<td>1.75</td>
<td>1.375</td>
<td>1.1875</td>
<td>1.0938</td>
<td>-</td>
</tr>
</tbody>
</table>

Obviously $R > 1$ when the marker restriction fragment is positively associated ($\Delta > 0$) with the disease, and vice versa. When $\Delta > 0$, the maximum value of $\Delta$ is given by $qM_{PD}$ if $P_M > P_D$ and by $qM_{PD}$ if $P_M < P_D$, so that we can write

$$R(\text{recessive}) = 1 + \frac{\Delta' q_M}{P_M} \quad \text{if} \quad P_M > P_D \quad [5a]$$

or

$$R(\text{recessive}) = 1 + \frac{\Delta' q_D}{P_D} \quad \text{if} \quad P_M < P_D \quad [5b]$$

where $\Delta' = \Delta / \Delta_{\max} (23)$, with $0 \leq \Delta' \leq 1$. Similar equations hold for the additive case, with the second term halved (see Eq. 4b).

Some representative values in the recessive case, with $\Delta'$ ranging from 1.0 to 0.0625, and $P_D$ values (or, equivalently, $P_M$ values) ranging from 0.01 to 0.4, are given in Table 1. The appropriate values for the additive case are obtained from the halved $\Delta'$ values in the recessive case; that is, for $\Delta' = 1$ in the additive case look up $\Delta' = 0.5$ in the recessive case.

If we require, for example, that the marker restriction fragment be at least 3 times more frequent in patients than in controls, then if $\Delta' = 1$ (the disease allele only occurs in association with this marker fragment) the method will detect the association for a recessive mode of inheritance of the disease if the control frequency of the marker restriction fragment is $<0.3333$. The frequency of the marker fragment in controls must be less than this if the association of the disease with the marker is not absolute ($\Delta' < 1$). The proportion of cases that will satisfy the conditions of Eqs. 5 (see Table 1) cannot be calculated, but since the map distance over which significant linkage disequilibrium is found in the HLA region has been estimated as 2% (W. Klitz, personal communication), it appears that this approach will have great utility.

When $\Delta < 0$, the maximum negative value that $\Delta$ can take is $-P_{PD}$ when $P_M + P_D = 1$ and is $-qM_{PD}$ when $P_M + P_D = 1$, so that when the marker restriction fragment is negatively associated with the disease ($R < 1$),

$$R(\text{recessive}) = 1 - \Delta' \quad \text{if} \quad P_M + P_D \leq 1 \quad [6a]$$

and

$$R(\text{recessive}) = 1 - \frac{\Delta' q_{M_{PD}}}{P_M P_D} \quad \text{if} \quad P_M + P_D \geq 1 \quad [6b]$$

with $0 \leq \Delta' \leq 1$. (The second term is again halved for the additive case.) In this case we should in fact look at the inverse of $R$, denoted $I$, since the marker fragment will be more frequent in controls than in patients. If we require that the marker restriction fragment be at least 3 times more frequent in controls than in patients, then negatively associated markers with $\Delta' > 0.6667$ will be detected.

N.A., C.S., and H.E. are grateful to Dory Bugawan for valuable technical assistance, to Michele Manos for helpful discussion, and to Carl Grumet, John Bell, Ronald LaPorte, Allan Drash, and Diane Wagener for providing HLA-DR-typed cells. We thank Elizabeth Jarvis for careful preparation of the manuscript. G.T. received support from National Institutes of Health Grant HD12731.