Review

Genetics of coeliac disease

R.S. HOULSTON and D. FORD
From the Section of Epidemiology, Institute of Cancer Research, Sutton, UK

Summary

Coeliac disease is one of the most common gastrointestinal disorders. The clinical features of the disease are protean, possibly due to heterogeneity. A familial basis for coeliac disease is well recognized, and although a strong HLA association is seen, this cannot entirely account for the increased risk seen in relatives of affected cases. A gene (or genes) at an HLA-unlinked locus also participates in causing coeliac disease and is likely to be a stronger determinant of disease susceptibility than the HLA locus. Such a gene (or genes) could theoretically act either additively or multiplicatively in conjunction with HLA. However, the familial risks seen in siblings and monozygotic twins are most parsimonious with a multiplicative model. Without evidence for a particular HLA-unlinked gene, and because no genetic model can be reliably ascribed to the non-HLA-linked locus, identifying causative non-linked HLA genes is likely to be through a genome-wide linkage search using non-parametric methods.

Introduction

Coeliac disease is a malabsorption disorder characterized by villous atrophy of the small-intestinal mucosa, caused by intolerance to gluten or related proteins in cereals such as wheat and rye. Coeliac disease occurs largely in Whites and rarely affects native Africans, Japanese or Chinese. Estimated prevalence rates in Europe range from a high of 1 in 300 in Western Ireland to between 1 in 1000 and 1 in 2000 in other regions. The true prevalence of coeliac disease in some studies is likely to have been underestimated, since a substantial number of individuals are asymptomatic or only have mild symptoms. Recent studies using antigliadin antibodies for screening have shown that the frequency of coeliac disease among some symptom-free individuals is high (1 in 256). Immunological factors are fundamental to the pathogenesis of coeliac disease, and current evidence suggests that the host immune responses to gliadin and related prolams are genetically determined. A familial basis for coeliac disease is well recognized. In addition to there being a strong HLA association, a gene (or genes) at an HLA-unlinked locus determines disease susceptibility.

Evidence for a familial predisposition to coeliac disease

Familial aggregation of coeliac disease was first reported over 60 years ago. Further studies of small-bowel biopsy specimens from first-degree relatives of patients with coeliac disease provided compelling evidence that genetic factors influence susceptibility to this disease. Different studies have reported varying risks of coeliac disease among first-degree relatives of patients. Reported estimates vary from under 5% to over 20%, with most ranging between 10 and 12%. This variation may result not only from genetic and environmental heterogeneity among populations, but from differing diagnostic criteria between studies. Additional support for an inherited predisposition to develop coeliac disease comes from twin studies.

Address correspondence to Dr R.S. Houlston, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG © Oxford University Press 1996
The concordance rate of coeliac disease seen in monozygotic twins is around 70%.\textsuperscript{16} Incomplete concordance between monozygotic twin pairs suggests that additional environmental factors are involved in the pathogenesis of coeliac disease. However, not all the twin pairs studied had proven monozygosity, and some of the twin pairs had insufficient long-term follow-up to be certain that the disease would not develop at a later stage.

**Association between coeliac disease and the HLA system**

Genes within the HLA gene complex exhibit a high degree of linkage disequilibrium, i.e. some alleles occur more often together on haplotypes (the particular combination of HLA alleles on a single chromosome) than expected from their respective frequencies. Linkage disequilibrium is primarily a function of (i) the physical distance separating the two loci, and (ii) the number of generations since the association first occurred in a founder. Other factors influencing linkage disequilibrium and which can be confounders in association studies include inbreeding, population admixture and mutation. Linkage disequilibrium can make it difficult to determine whether a particular gene is involved directly in disease susceptibility or whether it marks the effect of a linked gene.

Coeliac disease was first reported to be associated with the HLA class I molecule B8.\textsuperscript{17,18} Later, a stronger association to the class II molecule DR3 (now termed HLA-DR17) was found.\textsuperscript{19-21} However, the proportion of coeliac patients possessing HLA-DR17 varied considerably between populations (65%-95%), with a higher frequency among patients with a northern European ancestry. Other studies based on patients with a Mediterranean ancestry demonstrated an increase in DR11/DR7 haplotypes (DR11 and DR12 formerly termed DR5).\textsuperscript{22-24}

Subsequently, it became evident that the strongest association between HLA and coeliac disease was with the serological marker DQ2.\textsuperscript{25} The DQ2 molecule encoded by the alleles DQB1*0201 and DQA1*0501 is possessed by 95% of coeliac patients, compared to 20-30% of controls.\textsuperscript{25-30} In countries such as Japan where coeliac disease is uncommon, the frequency of the DQB1*0201 and DQA1*0501 haplotype is rare.

If serological methods are used to detect class II molecules, DQ2 is expressed in both DR17 and DR7 individuals.\textsuperscript{31,32} This is because serology only recognizes the β chain of DR2. It is now known that both the DQA and DQB genes coding for the DQ2 molecule are important for determining susceptibility to coeliac disease.\textsuperscript{33-35}

Sequence data on HLA-DQ genes has shown that because of linkage disequilibrium, the DQ2-DR17 haplotype carries the DQA1*0501 and DQB1*0201 alleles, the DQ2-DR11/DR12-haplotype carries the DQA1*0501 and DQB1*0301 alleles, and the DQ2-DR7 haplotype carries the DQA1*0201 and DQB1*0201 alleles.\textsuperscript{36-40} Although individuals with DR7-DQ2 haplotype carry a DQB1*0201 allele, they only have an increased risk of coeliac disease if they also carry DR11 or DR12. This is because the DQA1*0501, B1*0201) heterodimer that is encoded in cis (i.e. on the same chromosome), in DR17 individuals is encoded for in trans (i.e. on different chromosomes) by individuals with DR11/DR7 or DR12/DR7.

The possibility that the DQ2 molecule conferring susceptibility to coeliac disease might be unique has been suggested.\textsuperscript{35} This has been excluded by demonstrating that the DQB1*0201 and DQA1*0501 alleles, as well as other HLA class II alleles in coeliac disease, do not show any disease-specific sequences.\textsuperscript{33,41} Hence although there may be over-representation of specific class II molecules in coeliac disease these represent normal structural variants.

There is some suggestion of a gene dosage effect of DQB1*0201 from a study of Norwegian coeliac patients and controls selected for the presence of the DQA1*0501-DQB1*0201 haplotype. An excess of DQB1*0201 homozygosity was observed in patients compared with controls.\textsuperscript{30}

A very small proportion of coeliac disease patients (<5%) do not, however, possess the DQA1*0501, B1*0201) heterodimer,\textsuperscript{42,43} which may signify genetic heterogeneity. The majority of these cases possess DR4 and DQ8, the latter encoded by DQA1*0301 and DQB1*0302.

Although the principal association between coeliac disease and HLA is possession of the DQA1*0501/DQB1*0201 heterodimer, this does not preclude involvement of other genes in the HLA complex. The possibility that other class II genes are involved in the development of coeliac disease is suggested by the over representation of the rare DP alleles DPB1*0301 and DPB1*0101. The increase in frequency of DPB1*0301 in coeliac patients of North and South European ancestry appears independent of linkage to DQ2.\textsuperscript{33,44-47} However, the association of coeliac disease with DPB1*0101 can be ascribed to linkage disequilibrium in some populations.\textsuperscript{33,48,49}

Studies of other genes located between DP and DQ, including those involved in cellular processing such as TAP1 and TAP2 have not shown any association between polymorphic variation at these loci and coeliac disease.\textsuperscript{29,50,51}
Evidence for causative non-HLA genes in coeliac disease

The difference in concordance rates between monozygotic twins and HLA identical siblings (70% vs. 30%) implicates non-HLA genes in genetic predisposition to coeliac disease. This is further supported by the observation that coeliac disease occurs more frequently in family members who share the HLA susceptibility haplotype than in non-family members who appear to have the same HLA haplotype. One caveat of this is that studies which have been reliant on establishing HLA identity in the past depended largely on serological testing from markers at one or two HLA loci, and did not examine HLA identity across the entire HLA class D region. Thus individuals assumed to be HLA-class-II-identical may not be genotypically identical.

Using family data, Pena et al. proposed the involvement of two distinct unlinked genes in the aetiology of coeliac disease. A prerequisite for developing coeliac disease was homozygosity at the HLA-unlinked locus and participation of an independently inherited gene or genes located in or tightly linked to the major histocompatibility system acting in a dominant fashion. This proposal was supported by some but not all other studies. Most agreed on recessivity at the HLA-unlinked locus, but differed with respect to dominance or recessivity at the HLA-linked disease susceptibility locus. It is clear, however, that if expression of the DQ(α1*0501, β1*0201) heterodimer is responsible for the HLA association, constitutive haplotypes may not behave in a simple mendelian recessive or dominant fashion.

The second gene, unlinked to HLA, is likely to be a stronger determinant of disease susceptibility than the HLA-linked locus. The risk to an individual associated with DQB1*0201, DQA1*0501 or DQ(α1*0501, β1*0201) is no more than 50-fold increased over population risk. Based on a risk ratio of 50 and a carrier frequency of 0.34 for DQB1*0201, the relative risk of coeliac disease in siblings due to an HLA-linked locus would be 1.8 if the HLA-linked locus acted dominantly, or 4.6 if it acted as a recessive (Appendix 1 shows the calculation of relative risk from risk ratio and gene frequency). Alternatively a co-dominant model may be assumed to allow for a gene dosage effect of DQA1*0501-DQB1*0201. Given a carrier frequency of 0.24 for DQA1*0501-DQB1*0201, and risk ratios of 39 for heterozygosity and 308 for homozygosity, the relative risk of coeliac disease in siblings due to the HLA-linked locus is 4.0. Since the risk of coeliac disease in siblings is approximately 10%, the overall relative risk in siblings (due to HLA and other loci/environmental effects) is at least 20, and is therefore 4-fold higher than that attributable to HLA alone under any postulated model, assuming coeliac disease affects at most 1 in 200 individuals.

Identification of an HLA-unlinked gene causing coeliac disease

Identification of a non-HLA gene which predisposes to coeliac disease may eventually come through mutation testing of likely candidate genes in affected individuals, as such genes are characterized in the future. Alternatively, a genetic linkage search across families with multiple cases of coeliac disease could lead to the localization of a gene more quickly.

Classical linkage analysis is based on identifying the co-segregation of a DNA marker and disease in a family which cannot simply be ascribed to chance. The likelihood is a function of the chromosome distance separating the marker and the disease locus. Knowledge about the mode of inheritance of the disease is a prerequisite for undertaking the analysis.

If a genetic model of inheritance cannot be inferred, a non-parametric approach to linkage is adopted. This method is based on the typing of DNA markers in affected family members to identify a marker(s) where allele sharing between affecteds is significantly greater than expected. The ability of this type of approach to detect a disease-susceptibility locus depends on the contribution the locus makes to the genetic variation of the trait. This may be measured in terms of the increased risk to relatives of an affected individual as compared to the population prevalence (i.e. the relative risk).

A non-HLA-linked gene could theoretically act either additively (i.e. the penetrance of the disease is represented by the sum of the penetrances contributed by two or more loci) or multiplicatively (i.e. the penetrance of the disease is the product of the penetrances contributed by two or more loci) in conjunction with HLA. However, the familial risks seen in siblings (10%) and monozygotic twins (70%) are most parsimonious with a multiplicative model, since a simple additive model would violate the mathematical relationship which exists between sibling, parent-offspring and monozygotic twin relative risks. (The relative risk in monozygotic twins, λmtz, is given by: 4λs — 2λpo — 1, where λs and λpo are the relative risks in siblings and parent-offspring, respectively.)

Table 1 shows the overall relative risk of coeliac disease in first-degree relatives and the expected relative risks conferred by a putative non-HLA gene under a number of illustrative genetic models. Since the variable phenotypic expression of coeliac disease may reflect genetic heterogeneity, models based on
Table 1  Relative risk of coeliac disease in relatives of coeliac patients and risks conferred by non-HLA linked genes under a number of different genetic models

<table>
<thead>
<tr>
<th>Prevalence of coeliac disease</th>
<th>Overall</th>
<th>Xs</th>
<th>Amz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
<td>0.0033</td>
<td>0.002</td>
</tr>
<tr>
<td>(1/200)</td>
<td>(1/300)</td>
<td>(1/500)</td>
<td>(1/750)</td>
</tr>
</tbody>
</table>

- HLA-linked locus acting dominantly: $\lambda_s^H = 1.80$, $\lambda_{po}^H = 1.80$, $\lambda_{mz}^H = 2.60$
- Two equal risk genes acting multiplicatively with HLA
- Single gene acting multiplicatively with HLA
- Two equal risk genes acting multiplicatively with HLA
- Single gene acting multiplicatively with HLA
- Two equal risk genes acting multiplicatively with HLA

Based on a 10% sibling and a 70% monozygotic recurrence risk. The number of affected sibling pairs required to detect linkage under each model is also shown (see Appendix 2).

$\lambda_s$, sibling relative risk; $\lambda_{mz}$, monozygotic twin relative risk; $\lambda_{po}$, parent-offspring relative risk.

*Relative risks attributable to non-HLA-linked gene; **risks associated with the HLA-linked locus.

If coeliac disease can be ascribed to an abnormal T-cell-mediated response to exogenous antigens, genes outside the HLA system which could potentially contribute to disease susceptibility are those which influence this immune response. Although no association has been found between coeliac disease and T-cell-receptor polymorphisms or TAP1 alleles within the HLA-system, other genes involved in determining the T-cell immune response, such as the genes coding for the cytokines, cell adhesion...
molecules, etc. could be involved in causing coeliac disease. However, in the absence of good evidence for a particular HLA-unlinked gene and no reliable model for the inheritance of coeliac disease, the best approach for identifying causative non-linked HLA genes is likely to be through a genome-wide linkage search analysed using non-parametric methods.

References

Appendix 1

The sibling relative risk is defined as the risk of coeliac disease in a sibling of an affected proband compared with the risk in the population and may be computed as:

\[
\text{Sibling relative risk} = \frac{P(\text{affected sibling})}{P(\text{affected in the population})} = \frac{P(2 \text{ siblings affected})}{P(\text{affected in the population})^2}.
\]

Let \(A_0\), \(A_1\), and \(A_2\) denote the respective probabilities that two siblings are affected if neither, one, or both siblings are genetically susceptible to disease.

\[
A_0 = \binom{q_0}{2} + q_0 q_1 + q_1^2,
\]

\[
A_1 = q_0 (q_0 - 2q_1 + q_1^2) + 2q_0 q_1 - q_1^2,
\]

\[
A_2 = q_1 - 2q_0 + q_0^2 + q_1^2 - 1.
\]

and the population risk \(P(\text{affected in the population})^2\) is

\[
\left(\frac{1}{\mu} + \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)^2.
\]

The sibling relative risk is given by:

\[
\delta = \frac{1}{\mu} + \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2}.
\]

Appendix 2

Let \(\lambda_s\) denote the sibling relative risk, \(\lambda_{mz}\) the monozygotic twin relative risk, and \(\lambda_{po}\) the parent-offspring relative risk; where * refers to relative risks attributable to a non-HLA-linked gene and \(\lambda_s\) to the risks associated with the HLA-linked locus.
Under a multiplicative model, $\lambda s^*$ and $\lambda mz^*$ are estimated so that $\lambda s^* \times \lambda s^*$ (or $\lambda s^* \times \lambda s^* \times \lambda s^*$ for a model allowing for two non-HLA-linked genes) and $\lambda mz^* \times \lambda mz^*$ (or, similarly, $\lambda mz^* \times \lambda mz^* \times \lambda mz^*$) multiply to give the total sibling relative risk, $\lambda s$, and the total monozygotic twin relative risk, $\lambda mz$, respectively. $\lambda po^*$ are computed for any $\lambda s^*$, $\lambda mz^*$ by using the relationship:

$$\lambda mz^* = 4\lambda s^* - 2\lambda po^* - 1.$$  

[reference 58]

Probabilities of sharing 0, 1 and 2 alleles identical by descent at a linked marker are defined by: $Z_0 = 1/4\lambda s^*$, $Z_1 = (\lambda po^*)/2\lambda s^*$, $Z_2 = 1 - Z_0 - Z_1$, respectively. 58 Using these probabilities, the number of affected sibling pairs who share 0, 1, or 2 alleles out of a total of N affected sibling pairs can be simulated for given $\lambda s^*$ and $\lambda po^*$. The expected numbers of affected sibling pairs sharing 0, 1 or 2 alleles under no linkage are $1/4N$, $1/2N$ and $1/4N$. The difference between the observed and expected sharing distributions can be used to calculate a classical chi-squared on two degrees of freedom ($\chi^2$) without any genetic restriction. 59 A $\chi^2$ of 13.8 (equivalent to a LOD score of 3.0) was taken as statistically significant evidence for linkage. For a range of N, 10,000 simulations were performed and the percentage of simulations reaching statistical significance noted. The number of affected sibling pairs required to detect linkage given in Table 1 is the lowest N for which at least 90% of simulations achieved a $\chi^2$ of 13.8 or more.