Genetics in coeliac disease

David A. van Heel* DR
Karen Hunt
Department of Gastroenterology, Imperial College London, Du Cane Road, London W12 0NN, UK

Luigi Greco DR
Department of Paediatrics and International Laboratory for Food Induced Diseases, University of Naples Federico II, Naples, Italy

Cisca Wijmenga Professor
Complex Genetics Section, DBG-Dept. Medical Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands

Coeliac disease has a strong genetic component, higher than for many other common complex diseases. Possession of the HLA-DQ2 variant is required for presentation of disease causing dietary antigens to T cells, although this is also common in the healthy population. Non-HLA genetic factors account for the majority of heritable risk. Linkage studies have identified promising regions on chromosomes 5 and 19, with multiple other loci awaiting definitive confirmation in independent studies. Inherited variants in the tightly clustered chromosome 2q CD28-CTLA4-ICOS region are associated with disease, although of weak effect size. Larger sample sizes are necessary in coeliac disease genetic studies to detect small effects, alternatively meta-analysis offers promise. Newer methods including gene expression analysis and genome wide association studies will advance understanding of genetic susceptibility. Identification of coeliac disease genes may improve diagnostic/prognostic markers, basic understanding of disease aetiology, permit development of novel therapeutics and provide insight into other autoimmune disorders.

Key words: coeliac; genetic; linkage; genome scan; association; CTLA4; Celiac.

Coeliac disease is a chronic inflammatory disease of the small intestine induced by dietary proteins in wheat, rye and barley. Advances in understanding of disease immunology have identified immuno-dominant dietary (wheat gliadin) peptides resistant to intestinal enzymatic breakdown, modification of peptides by tissue transglutaminase (to which an antibody response is also made), and presentation of...
peptides to T cells by HLA-DQ2 as key steps leading to the intestinal inflammatory response. Although almost all Caucasians are exposed to potentially disease causing grains, present in substantial quantity in most Western diets, the prevalence of disease is \( \sim 1\% \).\(^1\) Twin and family based studies clearly show a strong genetic component to coeliac disease development, with inherited risk attributable to HLA and non-HLA factors (see Ref 23). Linkage studies have been performed to identify chromosomal regions likely to contain disease causing genes. Candidate gene association studies have focussed on genes known to be relevant to disease immuno-pathogenesis. This review will focus on attempts to identify inherited variants predisposing to coeliac disease and will discuss new developments. Identification of coeliac susceptibility genes may improve diagnostic and prognostic markers, allow better understanding of disease aetiology, permit development of novel therapeutics and resolve clinical overlap with other autoimmune disorders.

### GENETIC EPIDEMIOLOGY

Evidence that there is a strong inherited predisposition to coeliac disease susceptibility comes from twin studies and studies of prevalence in relatives of affected individuals. There is, however, no discernable Mendelian inheritance pattern in families, and risk falls more rapidly in distant relatives than would be expected in a disease caused by a single gene defect. Current theories suggest that multiple variants, each of relatively weak effect, may act together to influence disease risk in complex diseases.\(^2\)

Twin studies provide a powerful means of assessing the genetic and environmental components to disease susceptibility. Both monozygotic and dizygotic twin pairs share the same environmental factors, but differ by sharing 100 and 50\% of genetic variability, respectively. Twin studies depend on careful methodology because volunteer recruitment (e.g. through a patient group) may provide a selection bias towards monozygotic twins, and possible overestimation of the genetic component. Studies based on National Twin Registries, however, select twins independently of disease, which can be assessed separately. An early study, pooling data from 18 centres, found a 70\% coeliac disease concordance rate amongst monozygotic twins.\(^3\) A recent study used the large Italian twin registry of 1.6 million twins and identified index cases by cross-reference with lists of coeliac society members in Southern Italy.\(^4\) Non-index twins underwent comprehensive serological screening (with anti-endomysial or anti-tissue transglutaminase antibody), followed by small intestinal biopsy if positive. Fifteen out of 20 monozygotic twins were concordant for disease (75\%) compared to three out of 27 dizygotic pairs (11\%). The risk of disease in non-index monozygotic twins remained high after logistic regression analysis accounting for HLA haplotype sharing, sex and age, suggesting the involvement of non-HLA variants in disease susceptibility. By way of comparison monozygotic concordance rates are 25\% in multiple sclerosis,\(^5\) 36\% in type 1 diabetes\(^6\) and in 33\% in Crohn’s disease.\(^7\) Coeliac disease therefore appears to have one of the highest concordance rates of the complex multi-factorial diseases.

The sibling relative risk provides a further measure of the heritability of coeliac disease, defined as the risk to an affected patient’s sibling divided by the population risk (prevalence). Estimates in the Italian population\(^8\) suggest sibling relative risks of 48, although this might be an overestimate because the population prevalence was assessed without primary anti-endomysial or anti-tissue transglutaminase antibody screening.\(^1\) One British study estimated a sibling relative risk of 30 (using a population prevalence of 0.3\% and sibling
recurrence risk of 10%). These statistics also suggest a stronger genetic component to coeliac disease than many other complex diseases. Studies in these populations have estimated the proportion of sibling relative risk due to the HLA, in both published analyses the HLA genes contributed at most 40% of the sibling inherited risk. These data further confirm the major role of non-HLA genes in coeliac disease susceptibility.

THE HUMAN LEUKOCYTE ANTIGEN (HLA) COMPLEX

The HLA complex occupies a 4 Mb region on chromosome 6p21, and contains some 200 genes of which over half are known to have immunological function. Association of variants in the HLA complex with coeliac disease was first reported in 1972 using serological methods. Strong linkage disequilibrium occurs around the HLA region (that is variants some distance apart occur together more often than expected by chance), and the first reports of association were with B8 and DR3 alleles. These variants occur on the highly conserved A1-B8-DR3 extended HLA haplotype. Interestingly coeliacs are at substantially higher risk of type I diabetes, autoimmune thyroid disease, and other immunological conditions that are also associated with this haplotype. Later studies found that the strongest coeliac disease association was with HLA-DQ2 in particular with the combination of alleles encoding the alpha and beta chain variants DQA1*05 and DQB1*02 of the DQ2 heterodimer. The primary function of the DQ molecules is to present exogenous peptide antigens (e.g. in coeliac disease fragments of dietary wheat protein) to helper T cells. The numerous genetic polymorphisms found in HLA-DQ (e.g. DQ2) mainly alter amino acids in the peptide binding groove and affect the peptide binding/presentation repertoire. The genetic association data implicating DQ2 in coeliac disease is strengthened by functional studies identifying a specific immuno-dominant tissue transglutaminase modified wheat gliadin epitope recognised by DQ2 restricted T cells in the intestine and peripheral blood after antigen challenge, and the solution of the crystal structure of dominant epitope-DQ2 binding.

The HLA-DQ2 hetero-dimer can be encoded in cis (on the same haplotype), or more rarely in trans where the monomer subunits are encoded on separate haplotypes (Table 1). Although the cis and trans forms of DQ2 differ at one amino acid in each subunit, this is not thought to influence antigen presentation as the variants are in non-critical regions and both confer similar disease risk. Possession of the alpha chain DQA1*05 and beta chain DQB1*02 alleles (i.e. DQ2.5 in cis or trans) confers the primary increased risk of coeliac disease. Interestingly, however, several studies have shown that this risk is further increased by homozygosity for the cis haplotype or a second DQB1*02 allele on the other haplotype (i.e. DQ2.5/DQ2.5 or DQ2.5/DQ2.2. Table 1). This gene dosage effect has recently been shown in functional studies to be due to presentation of only a subset of peptides by the non-disease associated DQ2.2 molecule alone, and increased numbers of DQ2.5 dimers able to present antigen and stimulate effector T cells in the DQ2.5/DQ2.5 or DQ2.5/DQ2.2 states compared to simple DQ2.5 heterozygotes.

Around 90% of coeliac patients possess the HLA-DQ2.5 molecule, but more rarely other HLA class II molecules are associated with disease. Carriage of the HLA-DQ8 molecule (serologically DR4-DQ8 and at the genetic level DQA1*03-DQB1*0302) without DQ2 was found in 6% of coeliac cases in a recent large study of 1008 European coeliac patients. This study also found some evidence for a gene dosage effect of DQ8 similar to that for DQ2. Although gluten peptides presented by DQ8 have been
identified in studies using intestinal T cell clones, data confirming immunological function are still under scrutiny. Interestingly, in the European Study a further 6% of coeliac cases who did not carry classical DQ2.5 possessed either the alpha or beta chain of the heterodimer. Only four of 1008 apparent coeliac cases in the European cohort did not possess either DQ2 in part or DQ8.

Several studies have suggested that there may be additional alleles in the HLA region that confer increased disease risk over and above HLA-DQ. The high degree of polymorphism of the HLA and numerous potential candidate genes influencing immune function makes this an attractive hypothesis, and reported candidates include the TNF and MICA genes. Unfortunately, however, inadequate control for the strong linkage disequilibrium across the HLA is responsible for at least some of these claims. Two recent studies using larger cohorts and careful methodology did not find significant evidence for additional HLA risk factors, however, the question of HLA risk factors in addition to DQ2 remains unresolved.

The HLA-DQ2 allele is common in the healthy population, carried by approximately 30% of Caucasians. These data suggest that carriage of DQ2 is necessary but not sufficient for coeliac disease development, and combined with genetic epidemiological studies, have triggered the search for other non-HLA genetic variants predisposing to disease. There do not, however, appear to be genetic variants that exert a major influence similar to the HLA, it is likely that the large HLA effect size is related to the essential permissive role of DQ2 peptide presentation in disease pathogenesis.

### Table 1. Classical HLA DQ2 genotypes associated with coeliac disease, and gene dosage effects.

<table>
<thead>
<tr>
<th>Serological typing</th>
<th>Haplotype</th>
<th>DQ2 Genotype&lt;sup&gt;a&lt;/sup&gt;</th>
<th>DQ2 type (after&lt;sup&gt;17&lt;/sup&gt;)</th>
<th>Coeliac disease predisposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR3-DQ2/DR3-DQ2</td>
<td>1</td>
<td>DQB1<em>0201-DQA1</em>0501/</td>
<td>DQ2.5cis</td>
<td>Associated (high risk)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+DQB1<em>0201-DQA1</em>0501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR3-DQ2/DR7-DQ2</td>
<td>1</td>
<td>DQB1<em>0201-DQA1</em>0501/</td>
<td>DQ2.5cis</td>
<td>Associated (high risk)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ DQB1<em>0202-DQA1</em>0201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR3-DQ2/Other</td>
<td>1</td>
<td>DQB1<em>0201-DQA1</em>0501/</td>
<td>DQ2.5cis</td>
<td>Associated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR5-DQ7/DR7-DQ2</td>
<td>1</td>
<td>DQB1<em>0301-DQA1</em>0505/</td>
<td>DQ2.5trans</td>
<td>Associated&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+DQB1<em>0202-DQA1</em>0201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR7-DQ2/Other</td>
<td>1</td>
<td>DQB1<em>0202-DQA1</em>0201/</td>
<td>DQ2.2</td>
<td>Not associated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> DQA1*0501 and DQA1*0505 differ by one amino acid in the leader peptide, DQB1*0201 and DQB1*0202 differ by one amino acid in the membrane-proximal domain.

<sup>b</sup> This has been suggested to be an intermediate higher risk (vs. DQ2.5cis alone) group in the Southern European (French, Italian) populations in a recent study.

### GENOME WIDE LINKAGE STUDIES

Genome wide linkage studies aim to identify broad genomic regions which contain disease predisposing variants, and are a well-proven method to identify loci for
monogenic disorders (e.g. Cystic Fibrosis, Haemochromatosis). In monogenic disease, typically one or a few very large pedigrees are collected. Many parameters, such as mode of inheritance, penetrance and disease allele frequencies can be predicted from population analysis. This enables the construction of a genetic model which can be tested against the inheritance patterns observed using known polymorphic markers located throughout the genome. Linkage analysis in complex disease uses similar markers spread throughout the genome, however, because of the complex nature of inheritance it is difficult to specify a genetic model. In contrast to the parametric methods of monogenic disease, where the fit of a specific model is tested against the inheritance pattern of a disease causing gene, non parametric methods test whether the inheritance pattern amongst affected relatives differs from that expected by chance. For example, in the affected sibpair method, two affected siblings will share 0, 1 or 2 parental haplotypes with frequency 0.25, 0.5, 0.25, respectively (mean 0.5). Close to the disease locus, a chromosomal segment will be shared between the affected siblings and so sharing above 0.5 will be observed. These non-parametric methods are more robust, however, come at the cost of hundreds of families being necessary for adequate power.

In the more common polygenic diseases, the success of genome wide linkage studies has been limited until very recently. An appreciation of the need for tightly defined phenotypes and adequately sized cohorts has led to several disease genes being identified in the last few years, including the NOD2 gene for Crohn’s disease, ADAM33 for asthma, and phosphodiesterase 4D for ischaemic stroke. The genome scan approach has not yet led to the discovery of any disease predisposing variants for coeliac disease, although promising regions have been identified.

Eleven genome wide searches for linkage regions have been performed in coeliac disease, as well as numerous region specific linkage studies. Most studies have identified strong linkage to the HLA, providing a positive control, although this region is predictably easy to detect due to it’s large effect size. Individual genome scan results are shown in Table 2, and reported using the widely prevalent Lander and Kruglyak criteria for suggestive and significant linkage (expected by chance once per genome scan, and 1/20 genome scans, respectively). It is important, however, to note that application of these criteria, in the absence of study specific empirical simulations, may be too conservative for real data sets and that replication of weaker findings in independent cohorts can also be indicative of a true disease susceptibility locus.

Two loci are major interest, and likely to contain disease predisposing variants, on chromosome 5 and 19. One of the largest and most recent studies, performed in a Dutch cohort of coeliac patients meeting strict diagnostic criteria, identified significant linkage to chromosome 19p13.1. These authors also reported association of the microsatellite marker showing peak linkage (D19S899) in an independent case-control cohort using a multi-allelic test. The relevance of this statistic is uncertain, as association was not found with any one specific allele, which might be carried on a disease predisposing haplotype. Interestingly this region has also been strongly linked to Crohn’s disease, another inflammatory disease affecting the small intestine, in two independent genome scans suggesting a possible common predisposing variant.

A locus on chromosome 5q31–33 has been identified in Italian cohorts with some support from three other populations. A meta-analysis of 442 coeliac families provided convincing (but not independent) evidence for this locus (P = 0.000006). This region, which contains a cytokine gene cluster, has also been identified in asthma and Crohn’s disease genome scans. Furthermore a Crohn’s disease associated haplotype and putative disease causing mutation has been
Table 2. Coeliac disease susceptibility loci identified by genome-wide linkage studies (excluding the HLA).

<table>
<thead>
<tr>
<th>First author &amp; Year</th>
<th>Families in genome scan/ follow up studies</th>
<th>Population</th>
<th>Suggestive linkage(^a) ((P &lt; 7 \times 10^{-4}))</th>
<th>Significant linkage(^a) ((P &lt; 2 \times 10^{-6}))</th>
<th>Other Notable regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhong 1996(^39)</td>
<td>15</td>
<td>Irish</td>
<td>6p23, 11p11</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Greco 1998(^35, 2001)(^36)</td>
<td>39asp(^b)/71asp</td>
<td>Italian</td>
<td>5q</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Percopo 2003(^35)</td>
<td>89asp</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>King 2000(^49)</td>
<td>16</td>
<td>British</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>King 2001(^56)</td>
<td>34</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Naluai 2001(^38)</td>
<td>70/36</td>
<td>Scandinavian</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Liu 2002(^37)</td>
<td>60/38</td>
<td>Finnish</td>
<td>-</td>
<td>-</td>
<td>4p15</td>
</tr>
<tr>
<td>Woolley 2002(^57)</td>
<td>9/1 (large pedigree)</td>
<td>Finnish (Koilliskaira region)</td>
<td>-</td>
<td>15q12</td>
<td></td>
</tr>
<tr>
<td>Popat 2002(^56)</td>
<td>24</td>
<td>European</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Neuhausen 2002(^59)</td>
<td>62</td>
<td>North American</td>
<td>3p26, 5p14, 18q23</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>van Belzen 2003(^32)</td>
<td>67/15</td>
<td>Dutch</td>
<td>6q21–22</td>
<td>19p13.1</td>
<td></td>
</tr>
<tr>
<td>Rioux 2004(^41)</td>
<td>54</td>
<td>Finnish</td>
<td>10p</td>
<td>2q23–32</td>
<td></td>
</tr>
<tr>
<td>van Belzen 2004(^50)</td>
<td>1 (large pedigree)</td>
<td>Dutch</td>
<td>9p21–p13</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Reporting of loci according to strict Lander and Kruglyak criteria\(^29\) (only determined empirically in a single study\(^31\)).

\(^b\) asp: affected sibling pairs (family size not reported).
identified. However, no association of coeliac disease with the reported Crohn’s haplotype was found in two coeliac disease studies. Several studies in different populations have reported and replicated linkage to 2q33, although in only one genome wide scan did the evidence for linkage meet the strict suggestive criteria. This might reflect a weak effect size for this locus, hard to detect in the relatively small population samples available for study. The CTLA4 gene in the linked region has been investigated in multiple association studies, discussed in greater detail below.

At least 10 other loci have been reported with weaker significance from individual coeliac genome scans, and for these convincing replication data are awaited. The most promising is chromosome 9p21–p13 with suggestive evidence for linkage in the Dutch population and nominal evidence for linkage in the Scandinavian and Finnish populations. Chromosome 6q21–22 (distinct from the HLA) has been reported with suggestive linkage in one coeliac study and in type I diabetes, rheumatoid arthritis and multiple sclerosis. It is possible to speculate that a common variant at this locus might predispose to both coeliac disease and autoimmune diseases in general (as demonstrated by the HLA A1-B8-DR3 haplotype) and the non-random clustering of susceptibility loci in large numbers of genome scans for autoimmune disorders.

A large research effort has been expended in genome wide scans in coeliac disease, yet the true susceptibility loci seem hard to identify and there would appear to be a marked lack of consistency across studies. This picture is, however, common in other complex diseases and likely to represent weak statistical power, with loci unreported due to type I error. Studies have suggested that, for genes of moderate effect, sample sizes of less than 500 affected sibling pairs will give large variation in the magnitude and location of significant linkage results. Misclassification of disease status can also markedly weaken genome scan power, and heterogeneity might be reduced if studies used stringent criteria at diagnosis (the gold standard of crypt hyperplastic villous atrophy at intestinal biopsy, possibly combined with positive endomysial/anti-tissue transglutaminase serology). Further progress in coeliac disease gene identification by individual linkage studies appears unlikely without the use of larger cohorts. The European coeliac meta-analysis (four genome scans) has suggested a way forward, and might now be updated to include the several more recent datasets.

**CD28-CTLA4-ICOS VARIANTS (2Q33)**

The genomic region containing the CTLA4 gene has shown linkage to coeliac disease in several genome wide scans and replication studies (see above), although CTLA4 was initially studied as a candidate gene for coeliac disease based on known function. The T-lymphocyte regulatory genes CD28, CTLA4 and ICOS are found in a 300 kb block of chromosome 2q33. Much is now known regarding their immunological function: engagement of CD28 on naïve T cells by CD80/CD86 (B7) ligands on antigen presenting cells provides a potent co-stimulatory signal to T cells activated through their T cell receptor; engagement of ICOS on T cells by ICOS ligand (B7–H2) also provides a positive proliferative and cytokine secretion signal, although cell surface ICOS is expressed on activated rather than naïve T cells suggesting a later regulatory function; cell surface CTLA4 is also upregulated on T cell activation, has a higher affinity for its CD80/86 ligands than CD28, and provides a negative signal to regulate T cell activation. Hence all three genes control different aspects of the T cell response, and
**Table 3.** Association studies of the CD28/CTLA4/ICOS region in coeliac disease.

<table>
<thead>
<tr>
<th>First author &amp; Year</th>
<th>Population</th>
<th>Study design</th>
<th>Study size (families or cases/controls)</th>
<th>Number of markers tested a</th>
<th>Results for C49A/G</th>
<th>Results for other markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Djilali-Saiah 199860</td>
<td>French</td>
<td>Case/Control</td>
<td>101/130</td>
<td>1 SNP</td>
<td><em>P &lt; 0.0001</em> A allele</td>
<td>--</td>
</tr>
<tr>
<td>Holopainen 199966</td>
<td>Finnish</td>
<td>TDT</td>
<td>100</td>
<td>6 msat, 1 SNP</td>
<td>ns b D2S116 <em>P = 0.0001</em>, others + D2S2214 ns D2S2214 *P = 0.04 c, others + D2S1391, D2S116 ns D2S1391 *P = 0.03 c, D2S2214 *P = 0.007 c others ns</td>
<td></td>
</tr>
<tr>
<td>Clot 199971</td>
<td>Italian</td>
<td>TDT</td>
<td>192</td>
<td>1 SNP</td>
<td>ns</td>
<td>--</td>
</tr>
<tr>
<td>Tunisian</td>
<td>TDT</td>
<td>40</td>
<td>1 SNP</td>
<td>ns</td>
<td></td>
<td>--</td>
</tr>
<tr>
<td>Nalauai 200047</td>
<td>Scandinavian</td>
<td>TDT</td>
<td>107</td>
<td>5 msat, 1 SNP</td>
<td><em>P = 0.01</em> A allele D2S2214 *P = 0.04 c, others + D2S1391, D2S116 ns D2S1391 *P = 0.03 c, D2S2214 *P = 0.007 c others ns</td>
<td></td>
</tr>
<tr>
<td>King 200272</td>
<td>British</td>
<td>TDT</td>
<td>142 to 166</td>
<td>6 msat, 2 SNP</td>
<td>ns</td>
<td>--</td>
</tr>
<tr>
<td>King 200373</td>
<td>British</td>
<td>TDT + Case/Control</td>
<td>149 + 100</td>
<td>2 SNP</td>
<td>--</td>
<td>CTLA4_CT60 + other ns</td>
</tr>
<tr>
<td>Popat 200248</td>
<td>Northern European</td>
<td>TDT</td>
<td>116</td>
<td>7 msats, 1 SNP</td>
<td>ns d D2S2214 *P = 0.04 others + D2S1391, D2S116 ns</td>
<td></td>
</tr>
<tr>
<td>Martin-Pagola 200374</td>
<td>Basque</td>
<td>TDT</td>
<td>41</td>
<td>1 msat, 1 SNP</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Mora 200375</td>
<td>Italian</td>
<td>TDT + Case/Control</td>
<td>113</td>
<td>1 SNP</td>
<td><em>P = 0.03</em> A allele <em>P = 0.03</em> A allele</td>
<td>--</td>
</tr>
<tr>
<td>van Belzen 200366</td>
<td>Dutch</td>
<td>Case/Control</td>
<td>215/215</td>
<td>2 SNP</td>
<td>ns</td>
<td>CTLA4_CT60-G *P = 0.048</td>
</tr>
<tr>
<td>Rioux 200431</td>
<td>Finnish</td>
<td>TDT</td>
<td>54</td>
<td>1 SNP</td>
<td>ns</td>
<td>--</td>
</tr>
<tr>
<td>Haimila 200476</td>
<td>Finnish</td>
<td>TDT</td>
<td>106</td>
<td>8 msat, 17 SNP</td>
<td>ns</td>
<td>D2S2214.7 *P = 0.03, others + D2S116 ns CTLA4_CT60, ns ICOS_IVS+173T *P = 0.005 3 further ICOS SNPS *P = 0.03 CTLA4_+1822-T *P = 0.01 CTLA4_CT60-G *P = 0.04 CTLA4_haplotype_BI *P = 0.0007</td>
</tr>
<tr>
<td>Hunt 200564</td>
<td>British</td>
<td>Case/Control</td>
<td>340/705 to 973</td>
<td>7 SNP</td>
<td>ns</td>
<td>--</td>
</tr>
</tbody>
</table>

---

a snp: single nucleotide polymorphism; msat: microsatellite marker.

b ns, tested and non-significant (*P > 0.05*). Uncorrected results shown.

c Multi-allelic test.

d This snp was combined with further 58 families from a previous study77 in the analysis.
their close genetic proximity likely allowing for integrated control of expression. High expression of CTLA4 is also found on CD4^+ CD25^+ regulatory T cells, suggesting a further role for CTLA4 in suppressor T cell function.\textsuperscript{61,62}

Independent genetic association studies of the 2q33 region have now been performed by twelve groups (Table 3), using either case-control or family based methods. Case control cohorts are easier to collect (see Figure 1) and therefore can be larger with greater power than family based methods (e.g. TDT), although the latter are less prone to unrecognised population stratification. Some studies have focussed on multi-allelic microsatellite markers within the region, aiming to identify association with disease causing haplotypes rather than the markers per se. Association with coeliac

---

**CASE–CONTROL ASSOCIATION STUDY:** A study design in which cases with a defined condition and controls without this condition are sampled from the same population. Genetic factors (e.g. frequency of a SNP allele) are compared between the two groups to investigate the potential role of these in the aetiology. Population stratification may bias such studies.

**COMPLEX DISEASE:** A disease influenced by multiple environmental and genetic factors, and potentially by interactions in and between them.

**EFFECT SIZE:** The extent to which a factor influences the risk of the condition under study, rather than simply an indication of whether a factor is significantly related to the condition.

**GENOMIC CONTROL:** a method to assess population stratification by using data from a series of unlinked markers.

**HAPLOTYPE TAGGING:** The concept that most of the haplotype structure (combinations of alleles) in a particular chromosomal region can be captured by genotyping a smaller number of markers than all of those that constitute the haplotypes. The crucial ‘haplotype tagging’ markers are those that distinguish one haplotype from another\textsuperscript{90}.

**HUMAN LEUCOCYTE ANTIGEN (HLA):** 4Mb region of chromosome 6 containing many genes of immunological function.

**LINKAGE DISEQUILIBRIUM:** Two loci that are in linkage disequilibrium are inherited together more often than would be expected by chance (e.g. SNP variants close together on a chromosome). Occurs over small distances (~10-100kB) although highly variable.

**LINKAGE STUDY:** carried out in families containing multiply affected individuals to identify where disease genes are located. Polymorphic markers (e.g. microsatellites) are used to determine whether disease affected individuals share genetic information in a given region more often than expected by chance (because they also share a disease causing variant in the region). Linkage occurs over large distances around disease causing variants (e.g. ~10Mb).

**MICROSATELLITE:** A class of repetitive DNA sequences that are made up of organized repeats that are 2–8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetic studies.

**POPULATION STRATIFICATION:** occurs when a population consists of a set of (unrecognised) subpopulations (e.g. ethnic groups). If one subpopulation contains a frequency of disease allele that is relatively high, then any marker also at a higher frequency will appear to be associated, wherever it is located in the genome.

**POWER:** a measure of the probability that any given statistical test will detect a significant relationship when one actually exists in the data.

**SINGLE NUCLEOTIDE POLYMORPHISM (SNP):** Variation amongst individuals at a single base position in the genomic sequence e.g. a guanine (G) or thymidine (T). Usually limited to two alleles.

**TRANSMISSION DISEQUILIBRIUM TEST (TDT):** a method of detecting genetic association that avoids problems of population stratification. Instead of comparing unrelated cases and controls the test determines whether (given the parental genotypes) certain parental alleles are transmitted from parent to child more often than expected by chance.

---

**Figure 1.** Glossary of Genetic Terms.
disease has been found for the marker D2S2214, in four out of five studies genotyping this marker and with D2S116 and D2S1391 in single studies only. A meta-analysis of four studies also provided evidence for association at D2S2214 \((P = 0.001)\) and of three studies at D2S116 \((P = 0.0006)\) and D2S2392 \((P = 0.02)\). These markers are located 2–5 Mb from the CTLA4 gene, and show partial linkage disequilibrium with variants in the CD28/CTLA4/ICOS cluster. A more recent mega-analysis of raw genotype data showed weak association with three alleles of D2S2214 in 796 parent–child trios \((P < 0.05, < 0.01)\), and also with two other markers.

Most studies have analysed single nucleotide polymorphisms within the CD28/CTLA4/ICOS region. A French group first reported association of coeliac disease with the A allele at position +49 in the first exon of CTLA4 (amino acid changing, Ala17Thr) in a case-control study. Confirmatory results in independent populations have been demonstrated in only three of 13 follow-up studies which have genotyped this variant. A meta-analysis of eight studies provided modest evidence of association \((P = 0.02)\) with this variant using family based methods in 940 parent–child trios. Two of the largest studies have reported weak association with the recently described autoimmune disease associated functional CT60_G variant in the 3' region of the CTLA4, although two smaller studies have been negative. One recent study analysing all common SNP based CTLA4 haplotypes suggested the strongest association was at the haplotype rather than single variant level. Data on the function of these variants and haplotypes in coeliac disease will be necessary to determine whether the primary association is with variants in CTLA4, in a neighbouring gene or located over 2 Mb away (perhaps suggested by some of the microsatellite marker studies).

Exons of both CD28 and CTLA4 have been sequenced in large numbers of coeliac patients, and no evidence for mutations specific to coeliac disease has been found. Large scale studies in type I diabetes and autoimmune thyroid disease have recently clarified the haplotypic structure of the CD28-CTLA4-ICOS region, and a variant 6 kb 3' of the CTLA4 transcription start site (CT60) demonstrates the strongest autoimmune disease associations of over 100 variants studied from the region. This variant affects the mRNA levels of a secreted splice variant, sCTLA4, although the functional relevance remains unclear. The CTLA4+49 polymorphism creates an amino acid change in the leader peptide, and the +49G (Ala19) allele is associated with reduced glycosylation of the signal peptide, altered processing in the endoplasmic reticulum, and lower cell surface levels in transfected cells. The CTLA4 -318 promoter variant also appears to influence gene expression in reporter assays. Interestingly the NOD mouse, known to be predisposed to autoimmune disease, also has a defect in expression of a CTLA4 isoform, and recent data have suggested that the HLA-DQ8 transgenic mouse develops a coeliac disease/dermatitis herpetiformis like phenotype when crossed onto the NOD background.

**OTHER CANDIDATE GENES**

Many other candidate genes have been tested for association with coeliac disease based on knowledge of coeliac disease pathogenesis and immunology. Outside the HLA and 2q33 region, studies have analysed the FAS, MMP1/3, IL12b2, IRF1, DPPIV, TGM2, NOS2, KIR and LILR gene clusters and ELN genes. No convincing disease association has been found, although in most cases studies have not been powered to detect small effect sizes (see below). An interesting report of association with variants in the interferon
gamma gene\textsuperscript{78} was not confirmed in a larger Dutch sample.\textsuperscript{79} Further data from linkage and gene expression studies is necessary to identify novel candidate genes and allow prioritisation for future studies.

Data showing lack of replication of reported association, or lack of association with SNPs in strong candidate genes, are often difficult to publish and the studies above probably represent only a fraction of those performed. Publication, or a database, of well conducted important negative studies would allow researchers to focus on unexplored areas. Association can be either direct (the disease-causing variant itself), or indirect (via linkage disequilibrium) and lack of knowledge of linkage disequilibrium structure might also be responsible in part for inconsistencies between populations. Meta-analysis of candidate gene association studies is a promising and powerful tool, and can provide evidence for unexpected diversity in studies of similar populations as well as publication bias.\textsuperscript{80} It is not, however, a replacement for adequately powered original studies.

**GENE EXPRESSION PROFILING**

Microarray technology now allows expression (mRNA) levels in many thousands of genes to be assessed simultaneously, with near full coverage of all genes in the human genome beginning to be approached by some companies (e.g. Affymetrix Genechips covering 38,500 characterised genes). It has been hoped that an insight into genes containing primary disease causing functional alterations might be identified from such analyses, which can then be analysed in targeted genetic association studies to identify disease causing polymorphisms. Although there are a few notable exceptions,\textsuperscript{81,82} success of this approach in complex disease has been limited.\textsuperscript{83} Some of the potential pitfalls include identification of innocent bystander genes (whose expression levels have been altered by primary changes elsewhere); failure to detect genes that lead to small (e.g. two-fold) changes in expression but have profound effects on phenotype\textsuperscript{83}; not identifying amino acid changes that do not affect expression; and dilution of the cell type showing transcriptional changes in a complex whole tissue sample or conversely failing to analyse the cell expressing disease causing genes in a homogenous preparation (prepared by purification or laser capture microdissection). Furthermore gene function usually occurs at the protein level with an intermediary mRNA step, and gene function is often also regulated at a post-transcriptional level rather than solely at the mRNA level (a good example being TNF\textsubscript{α}).

In coeliac disease, two microarray studies on whole biopsy samples have been published.\textsuperscript{84,85} One Finnish study used a 5000 gene array (note there are ~30,000 human genes) to compare duodenal biopsies from untreated coeliacs, gluten free diet treated coeliacs and healthy controls.\textsuperscript{85} Use of histologically abnormal samples will identify changes due to inflammation (including cell recruitment), but is likely to obscure primary disease causing alterations. In the comparison between histologically normal treated coeliac and control mucosa (using a small sample size of $n = 4$ each), sixty genes showed changes of $> 1.25$ in expression levels in all sample pairs. No genes from candidate regions on 2q, 5q or 15q (see above) showed consistent expression changes, although only 12% of genes from these regions were analysed on the arrays. A second study compared coeliac biopsies from 15 patients with villous atrophy to seven normal control samples,\textsuperscript{84} using more comprehensive arrays containing 20,000 human genes. Genes and pathways were identified that might be important in the pathogenesis of enteropathy (characterised by crypt hyperplasia and villous atrophy).
The main aim of this study was to obtain insight into disease pathogenesis, which is currently limited to mechanisms of inflammation. The changes identified are unlikely to be primary disease causing, although may highlight candidate genes for further genetic studies. The preliminary Finnish data suggest a larger study using increased numbers of individuals in a comparison between histologically normal treated coeliac and controls in an attempt to identify causal changes. Careful attention to normal histology to minimise effects solely due to inflammation (including quantitative studies), use of increasingly sophisticated bioinformatics techniques of data analysis as well as whole genome expression arrays will maximise chances of successful candidate gene identification.

**DEVELOPMENTS IN GENETIC ANALYSIS**

Our current knowledge of coeliac disease susceptibility loci suggests focused research efforts on identifying disease causing genes from the most promising linkage regions, including 5q and 19p. Incorporation of data from expression analyses, protein interaction studies, transcription factor binding sites, large scale RNA interference studies and progress in gene annotation databases will help prioritise candidate genes.

Coeliac disease has a high heritable component to disease susceptibility and identification of disease causing variants should be achievable. It is expected that genes of large effect (e.g. HLA-DQ2) will be few and that genetic risk will be conferred by multiple genetic variants of individual weak effect. Linkage studies will require 1000's of families to approach adequate statistical power to identify such regions, which may not be realistically achievable. An alternative strategy is to perform a genome wide search for association, which is now becoming possible with improvements in genotyping technology. These studies will need to analyse ~200,000 haplotype tagging SNPs, and because of the relatively weak effect sizes expected and large numbers of statistical tests require large and multiple cohorts. Data from the International HapMap Consortium will allow a minimal SNP coverage to identify common haplotype blocks. The first large scale SNP based genome wide association study has recently been completed, using 66,000 variants, and led to identification of a novel predisposing variant for myocardial infarction.

The field of coeliac disease genetics would also be well served by attention to careful design of association studies. Case-control studies in epidemiology attempt to maximise success by sampling cases from the tails of phenotype distribution. This might be achieved in coeliac disease, which is fortunate in having readily available and clearly defined biological markers, by selection of early onset, DQ2 positive symptomatic patients with positive serology and villous atrophy on diagnostic biopsy. Such a strategy may also improve power by reducing heterogeneity. Use of much larger numbers of patients and controls, or parent–child trios, is also necessary to identify genes of weak effect. Studies that may be adequately powered to detect the HLA (a large effect size) are not suited to detection of the more common low risk variants predicted to underly complex disease. For example identification of a rare disease allele (5% frequency) at 80% power, \( P < 0.001 \), with odds ratio (OR) 1.3 requires 10,000 cases/controls, OR 2.0 requires 1100 cases/controls. A more common allele (20%) with OR 1.3 would require 2900 cases/controls and OR 2.0 require 360 cases/controls. Poor statistical power probably underlies much of the variation in association results across studies. Studies in other complex diseases, such as type I diabetes, are starting to achieve these large sample
sizes. Acquisition of these cohort sizes may require a move towards case-control association studies rather than use of family based designs such as the transmission disequilibrium test. Genomic control methods to assess and correct for undetected population structure (e.g. ethnic group mismatching) in case-control studies are now available and appear promising, although this may be a more minor issue than initially supposed. Attention is also needed to the reporting of positive associations. In addition to adequate sample size, an attempt to replicate primary findings in a second independent sample is critical, as well as a check for Hardy Weinberg equilibrium in the control cohort and attention to multiple testing of variants and phenotypes. The problems of reporting genetic associations in complex disease are well recognised.

Gene–gene interactions (epistasis) and gene–environment interactions are likely to contribute to risk of disease as well as additive effects from multiple disease causing variants. Epidemiological strategies and advances in statistical tools will be needed to define such effects.

CONCLUSIONS

Coeliac disease is amongst the most heritable of the common complex diseases, and should be amenable to identification of genetic variants predisposing to disease development. Substantial progress has been made in identifying broad chromosomal regions containing susceptibility loci. A major gene of strong effect (HLA-DQ2) is well understood in terms of both genetic and functional effects, although this accounts for a minority of disease heritability. Disease predisposing variants of much weaker effect have been identified from the CD28/CTLA4/ICOS gene cluster. Although difficult to identify, primary disease causing polymorphisms are emerging for other complex diseases (e.g. Crohn’s disease, type I diabetes, asthma) and promising new strategies (e.g. genome wide association studies) being developed. Newly identified coeliac disease susceptibility genes may lead to better diagnostic tools, novel targets for therapeutic intervention and improved understanding of autoimmune disease in general.

Research agenda

- Genome wide linkage searches have identified some coeliac disease susceptibility loci, although these are relatively poorly defined and poorly replicated. Much larger studies would be needed for comprehensive identification, although this strategy may be superceded by genome wide association studies (see below)
- Comprehensive SNP based linkage disequilibrium mapping studies of the 2q locus in large cohorts are necessary to identify the primary coeliac disease causing variant (suggested to be in the CD28/CTLA4/ICOS gene cluster)
- Linkage disequilibrium mapping studies of other promising loci (e.g. 5q, 19p) are needed to identify other disease genes
- Focused association studies of variants in candidate genes in known and novel (identified by microarray analysis) pathways implicated in disease pathogenesis may prove successful. Such studies need multiple cohorts and functional data to provide convincing evidence of causality
- Genome wide association studies in multiple large cohorts offer promise to identify disease causing variants of weak effect, and are just becoming feasible with current technology
ACKNOWLEDGEMENTS

David van Heel is funded by a Wellcome Trust Clinician Scientist Fellowship. Cisca Wijmenga is funded by the Netherlands Organization of Scientific Research, the Dutch Digestive Disease Foundation and the Celiac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009). Luigi Greco is funded by the Ministry of University and Research (MIUR grant PRIN 2003).

REFERENCES


