ARTICLE

Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21

Mathias Herr, Frank Dudbridge, Patrizia Zavattari1, Francesco Cucca1, Cristian Guja+, Ruth March2§, R. Duncan Campbell2¶, Anthony H. Barnett3, Stephen C. Bain3, John A. Todd‡ and Bobby P.C. Koeleman†

Wellcome Trust Centre of Molecular Mechanisms in Disease, Box 139, Department of Medical Genetics, University of Cambridge, Hills Road, Cambridge CB2 2XY, UK, 1Dipartimento di Scienze Biomediche e Biotecnologie, Universita’ di Cagliari, Cagliari, Italy, 2MRC Immunochemistry Unit, University of Oxford, Oxford OX1 3QU, UK and 3Department of Medicine, University of Birmingham, Birmingham B9 5SS, UK

Received 7 March 2000; Revised and Accepted 23 March 2000

The positional cloning of multifactorial disease genes is a major challenge in human genetics. We have therefore empirically tested the utility of the available polymorphic microsatellite map to locate the already identified type 1 diabetes locus IDDM1 (sibling risk/population prevalence ratio \( \lambda_s = 2.7 \)) within a 14 Mb region of chromosome 6p21 linked to disease. In a two-stage approach to fine mapping, linkage was evaluated in 385 affected sib-pair families using 13 evenly spaced polymorphic microsatellite markers. The whole 14 Mb showed strong linkage. Then, each marker was analysed for evidence of allelic association, revealing evidence of disease association at one marker located within the 95% confidence interval of 1.7 cM obtained by linkage. Analysis of an additional 12 markers flanking this marker revealed a highly specific region of 570 kb associated with disease (\( P = 7.5 \times 10^{-35} \)), which included the HLA class II genes, known to be the primary determinants of IDDM1. The peak of association was as close as 85 kb centromeric of the disease-predisposing class II gene HLA-DQB1. We investigated the importance of the underlying inter-marker linkage disequilibrium, marker informativity and recombination for fine mapping and demonstrate that the majority of disease association in the region can be explained by linkage disequilibrium with the class II susceptibility genes. Recombination within the major histocompatibility complex was rare and nearly absent in the class III region. We demonstrate that fine mapping of a multifactorial disease gene is possible with high accuracy even in a region with extraordinary linkage disequilibrium across distances of several Mb. The results will be applicable to association studies of disease loci with \( \lambda_s \) values <2.7 except that much larger data sets will be required.

INTRODUCTION

A major challenge in human genetics is the positional cloning of multifactorial disease genes. However, any attempt to map a gene predisposing to a multifactorial disease is complicated by phenomena such as locus and allelic heterogeneity, incomplete penetrance, and the fact that a combination of multiple genetic and environmental factors may influence the disease risk with each factor making a modest contribution (1–3). Even in a carefully ascertained family collection with precisely defined phenotypes, these factors are difficult to eliminate and may hamper the fine mapping of a complex disease gene.

Linkage analysis, as a first step in positional cloning of a multifactorial disease locus, is not expected to narrow down a candidate region to <1 Mb (4,5). In contrast, association...
Markers were chosen aiming for a heterozygosity and an average physical inter-marker distance of 1.3 Mb. Using a detection map of 13 physically even-spaced markers, descent (IBD) sharing analysis of affected sib-pair families was performed by identity by linkage analysis. Multipoint linkage analysis was initially performed on the same detection map as used for linkage analysis. We applied two multiallelic association analysis (bottom) with $T_{sp}$ (solid line) and ETDT (dotted line) were performed on a detection map with 13 physically even-spaced markers and an average physical inter-marker distance of 1.3 Mb. The total physical length of 14.1 Mb applied to a genetic length of 15 cM. The linkage 95% confidence interval (CI) of ~1.7 cM is defined by the two points on either side of the maximum LOD score, at which the LOD score is within the threshold $T_{y}$ (here $T_{y} = 1.47$) which was calculated according to Kruglyak and Lander. Orientation of the map is from centromere (left) to telomere (right).

In summary, LD mapping results may be characterized by tremendous stochastic variation, making the localization of a disease locus to a small interval potentially very difficult. In this study, we applied the strategies of linkage and association mapping to the known major susceptibility locus (IDDM1) of type 1 diabetes as a prime example of a multifactorial disease in order to test how accurately this known locus could be mapped. A total of 385 affected sib-pair families from the UK were genotyped for 25 polymorphic markers in a 14 Mb region on chromosome 6p21 including the human leukocyte antigen complex (HLA). The HLA class II genes are likely to be the main determinants of the type 1 diabetes locus mapped to this region, designated IDDM1 (12–14). Here we evaluate the influence of marker informativity, inter-marker LD, LD between markers and disease loci and recombination on association mapping and discuss the impact of single-point and two-point marker haplotype analysis. Furthermore, microsatellite markers are tested for independent association corrected for LD with the main susceptibility loci DQB1 and DRB1. Given our concerns about the non-linearity of LD, we were surprised to observe a very clear single peak of association, directly at the HLA class II loci. Our results provide the clearest graphic demonstration yet that, at least in the UK population, the HLA class II genes are likely to be the major contributors to IDDM1.

RESULTS

Linkage analysis

Multipoint linkage analysis was performed by identity by descent (IBD) sharing analysis of affected sib-pair families using a detection map of 13 physically even-spaced markers and an average physical inter-marker distance of 1.3 Mb. Markers were chosen aiming for a heterozygosity ≥60%. The total genetic length of the 14.1 Mb region was 15 cM. The average genetic inter-marker distance, calculated from our family set, was 1.25 cM. The individual distances between the markers, however, were different from the known physical distance, with overall smaller genetic intervals in the HLA region and larger intervals in the flanking region (Fig. 1). Linkage analysis, based on the genetic map, showed highly significant linkage to type 1 diabetes over the whole region with a minimum LOD score of 24.8 at the centromeric end (D6S1641) and a maximum LOD score of 43.3 between D6S1560 and D3A (Fig. 1). The 95% confidence interval calculated according to Kruglyak and Lander extended over an area of ~1.7 cM, marked by D6S1560 and D3A. The 95% confidence interval was calculated according to Kruglyak and Lander (5) extended between C1-3-2 and D6S265, which applies to a physical distance of ~2.7 Mb. This interval included the HLA class II and centromeric parts of the class I region.

Linkage disequilibrium fine mapping

Association analysis was performed on the same detection map as used for linkage analysis. We applied two multiallelic
association tests, \( T_{sp} \) (15) and the extended transmission disequilibrium test (ETDT) (16), which are based on the transmission disequilibrium test (TDT) (17) and correct for linkage within sib-pair families. With both \( T_{sp} \) and ETDT we identified an \(-2.2 \text{Mb} \) candidate region, marked by the microsatellites D6S1560 and C1-3-2, which was located within the 95% confidence interval previously obtained by linkage analysis. As shown in Figure 1, peak association was seen with marker D3A in both tests \([T_{sp}; \text{log10}(P) = 20.5; \text{ETDT}; \text{log10}(P) = 18.3]\).

For fine mapping we focused on the \(2.2\text{ Mb}\) candidate region obtained by both linkage and association analysis and typed an additional set of 12 microsatellite markers available at that time within this interval. Association with disease was now much stronger and most significant at markers in the HLA class II complex over a region of only 570 kb, flanked by markers D6S2444 and D3A. As shown in Figure 2, the peak of association had shifted to marker D6S2444 in both the \( T_{sp} \) and ETDT single-point analysis \([T_{sp}; \text{log10}(P) = 33.1; \text{ETDT}; \text{log10}(P) = 28.1]\). This microsatellite is located 85 kb centromeric of HLA-DQB1 and 160 kb centromeric of HLA-DRB1, and accordingly is the closest marker to the main susceptibility genes of those typed in this study. The disease association of these class II genes themselves was as follows: HLA-DQB1, \(-\text{log10}(P) = 64.7\) (\( T_{sp} \)) and \(-\text{log10}(P) = 66.9\) (ETDT); HLA-DRB1, \(-\text{log10}(P) = 59.0\) (\( T_{sp} \)) and \(-\text{log10}(P) = 58.2\) (ETDT).

Association was considerably weaker, but still highly significant throughout the HLA class III region with peaks at markers 82-1 and C1-2-C, and fell off \(-1.5 \text{Mb}\) telomeric of the DQB1 locus at marker C1-3-2 (140 kb telomeric of HLA-C). Markers 9N-1 and D6S273 showed a considerable drop of association within this highly associated region (see below). Between C1-3-2 and D6S2223 association remained on a much lower significance level around \( P = 0.01 \). In contrast to the telomeric side, the whole \(7.6 \text{Mb}\) region centromeric of D6S2445 showed no significant association with the disease.

Association tests were also performed on two-marker haplotypes instead of using single markers. Peak association was seen with the D6S2444–D3A haplotype (Table 1). Analysing two-marker haplotypes improved to some extent the informativity of certain markers, thus leading to stronger association; for instance, 698 parents were heterozygous for the D6S2444–D3A haplotype compared with 578 for D6S2444 and 532 for D3A on their own. Furthermore, the above-mentioned drop of single-point association for markers 9N-1 and D6S273 was alleviated when both these markers were tested as a two-marker haplotype (Table 1). In this case, while the heterozygosity of each single marker and the corresponding haplotype was similar, the stronger association of the two-point haplotype was due to an increase of LD with the major disease locus DQB1–DRB1, as further discussed below.

There was little difference between both \( T_{sp} \) and ETDT concerning the accuracy of mapping and significance of association at each marker. All single-point and two-point haplotype association results for the microsatellite markers are listed in Table 1.

To establish an empirical confidence interval around the peak of association we performed a percentile bootstrap procedure \((18)\). A total of 385 families were selected randomly with replacement from the 385 families, which means that some families can occur more than once in one such replicate. The procedure was repeated 1000 times. In each of the 1000 replicates the peak of association was established, and the empirical confidence interval was calculated from the number of times a marker gave the peak of association. Since in these calculations the maximal peak of association was found more often at the marker closest to DQB1–DRB1 with \( T_{sp} \) than with ETDT, results are given as follows for \( T_{sp} \) (ETDT in parentheses). In 98.2% (93.6%) of the replicates the maximal \(-\text{log10}(P)\) was found at marker D6S2444. Of the remaining 1.8% (6.4%) the peak was detected in 1.5% (3.8%) at marker D3A and in 0.1% (1.4%) at 82-1. The D6S2444–D3A haplotype was found to be the most associated two-point haplotype in 99.8% (98.9%) of the replicates. These data show that maximum association could be mapped with very high confidence to marker D6S2444.

Linkage disequilibrium Distribution of LD over the region was studied using the total disequilibrium measure \( D' \). This measure is illustrated for gametic haplotypes with the corresponding \( P \)-value and for each marker pair related to the physical distance (Fig. 3). It demonstrates a clear drop of the total LD (above) and the significance of LD (below) centromeric of D6S2444 (~85 kb centromeric of DQB1). Moreover, the graph illustrates the extending LD from the HLA class II region to the telomeric side over a distance as far as 5 Mb. Even between equally spaced markers, the magnitude of LD was greater in the HLA region (e.g. between D6S265 and D6S258; inter-marker distance 1.25 Mb; \( D' = 0.35; P \leq 1 \times 10^{-7} \)) than in the region centromeric of the HLA complex (e.g. between marker D6S1576 and D6S291; inter-marker distance 1.4 Mb; \( D' = 0.089; P = 0.29 \)). In addition, we investigated the total LD in affected family-based control (AFBAC) chromosomes (19). The result (data not shown) was very similar to that illustrated in Figure 3.

We also evaluated LD in the context of the known disease genes DQB1 and DRB1. Figure 4 shows the total \( D' \) calculated...
Table 1. Microsatellite marker characteristics and association data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Position</td>
<td>0</td>
<td>1.5</td>
<td>2.9</td>
<td>3.5</td>
<td>5.9</td>
<td>6.1</td>
<td>6.9</td>
<td>7.1</td>
<td>7.5</td>
<td>7.7</td>
<td>7.9</td>
<td>8.3</td>
<td>8.6</td>
<td>8.9</td>
<td>8.9</td>
<td>8.1</td>
<td>9.3</td>
<td>9.6</td>
<td>9.8</td>
<td>9.9</td>
</tr>
<tr>
<td>Number of</td>
<td>13</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>12</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>11</td>
<td>21</td>
<td>13</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>alleles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>alleles w/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequency  &gt;0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td>605</td>
<td>674</td>
<td>543</td>
<td>709</td>
<td>678</td>
<td>604</td>
<td>649</td>
<td>429</td>
<td>657</td>
<td>740</td>
<td>721</td>
<td>746</td>
<td>707</td>
<td>708</td>
<td>692</td>
<td>720</td>
<td>656</td>
<td>618</td>
<td>560</td>
<td>660</td>
</tr>
<tr>
<td>fully typed</td>
<td>89.0</td>
<td>87.5</td>
<td>70.5</td>
<td>92.0</td>
<td>88.1</td>
<td>78.4</td>
<td>84.3</td>
<td>55.7</td>
<td>85.3</td>
<td>96.1</td>
<td>93.6</td>
<td>96.9</td>
<td>91.8</td>
<td>91.4</td>
<td>89.9</td>
<td>93.5</td>
<td>85.2</td>
<td>80.3</td>
<td>73.9</td>
<td>85.7</td>
</tr>
<tr>
<td>parents (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of</td>
<td>490</td>
<td>410</td>
<td>231</td>
<td>510</td>
<td>454</td>
<td>452</td>
<td>578</td>
<td>303</td>
<td>378</td>
<td>578</td>
<td>532</td>
<td>538</td>
<td>569</td>
<td>574</td>
<td>575</td>
<td>616</td>
<td>525</td>
<td>531</td>
<td>438</td>
<td>498</td>
</tr>
<tr>
<td>heterozygous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parents (%)</td>
<td>71.5</td>
<td>60.8</td>
<td>42.5</td>
<td>71.9</td>
<td>67.0</td>
<td>74.8</td>
<td>89.1</td>
<td>84.6</td>
<td>57.5</td>
<td>78.1</td>
<td>73.8</td>
<td>72.1</td>
<td>80.5</td>
<td>81.1</td>
<td>83.1</td>
<td>85.6</td>
<td>80.0</td>
<td>83.9</td>
<td>77.0</td>
<td>75.5</td>
</tr>
<tr>
<td>Number of</td>
<td>500</td>
<td>356</td>
<td>382</td>
<td>507</td>
<td>449</td>
<td>472</td>
<td>333</td>
<td>314</td>
<td>485</td>
<td>698</td>
<td>555</td>
<td>557</td>
<td>533</td>
<td>552</td>
<td>547</td>
<td>536</td>
<td>442</td>
<td>393</td>
<td>421</td>
<td>457</td>
</tr>
<tr>
<td>heterozygous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parents (%)</td>
<td>50.0</td>
<td>35.6</td>
<td>38.2</td>
<td>50.7</td>
<td>44.9</td>
<td>47.2</td>
<td>33.3</td>
<td>31.4</td>
<td>48.5</td>
<td>69.8</td>
<td>55.5</td>
<td>55.7</td>
<td>53.3</td>
<td>55.2</td>
<td>54.7</td>
<td>53.6</td>
<td>44.2</td>
<td>39.3</td>
<td>42.1</td>
<td>45.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Association for single-point markers</th>
<th>Tsp (-\log(P))</th>
<th>ETDT (-\log(P))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>1.05</td>
<td>1.3</td>
</tr>
<tr>
<td>ETDT (-\log(P))</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Tsp (-\log(P))</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>ETDT (-\log(P))</td>
<td>0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Association for two-point marker haplotypes</th>
<th>Tsp (-\log(P))</th>
<th>ETDT (-\log(P))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>0.47</td>
<td>0.99</td>
</tr>
<tr>
<td>ETDT (-\log(P))</td>
<td>0.31</td>
<td>0.38</td>
</tr>
</tbody>
</table>

\*Figures for two-point marker haplotypes reflect the marker in the respective column combined with the marker in the following column of the table.
between each marker and DQB1–DRB1. Furthermore, since total D‘ values may not be informative for strong haplotype-specific LD, pairwise D‘ values were calculated between both the main disease haplotypes (DQB1*0302–DRB1*0401 and DQB1*02–DRB1*03) and the most frequent allele and two-point haplotype of each microsatellite marker on these two haplotypes (Fig. 5). Both analyses demonstrate in more detail that LD with the major disease loci is strong on the telomeric and weak on the centromeric side and illustrate how the pattern of association reflects the pattern of LD.

Conditional ETDT (CETDT)

We used the CETDT (20) to test 21 markers located within and adjacent to the HLA region for an independent susceptibility effect corrected for LD with DQB1–DRB1. Briefly, the CETDT estimates transmission probabilities of all DQB1–DRB1–marker haplotypes. Then a likelihood ratio test is performed to test whether haplotypes with identical alleles at DQB1–DRB1 but different alleles at the marker have identical transmission probabilities. D6S1568, positioned 300 kb centromeric of DPB1 and 80 kb centromeric of the Tapasin gene, showed a trend to independent association (P = 0.052) which was detected also when conditioning on only DQB1*02–DRB1*03 (DR3) (P = 0.055), but not when conditioning on DQB1*0302–DRB1*0401 (DR4). The only marker showing significant independent disease association was D6S2223 (P = 0.0018; after Bonferroni correction: Pc = 0.038). This marker is located 5.9 Mb telomeric of the DRB1 locus. D6S2223 also showed an independent effect on the DR3 haplotype (P = 0.003) but not on the DR4 haplotype. In the ETDT analysis the DQB1*02–DRB1*03–D6S2223 haplotype had a lower transmission weight when carrying allele 3 (181 bp) at marker D6S2223 than when carrying any of the other alleles.

A different method to correct the association of a tested marker for its LD with another associated locus/marker is the homozygous parent TDT (21). In this test D6S1568 showed some increased transmission of allele 8 (105 bp) on the DR3 haplotype, but the transmission counts were very low (10 transmitted, three not transmitted; exact P = 0.09). For D6S2223 transmission counts were too low to detect an independent effect.
All families were investigated for recombination using the sibmap function of the GAS software, which assesses crossovers by changes in allele-sharing status between siblings given a certain marker order. Recombination was considered only when marked by at least two informative markers on either side of the recombinant chromosome interval. In total, we observed 118 recombinant chromosomes, 79 maternal and 39 paternal, which equals a female:male ratio of 2.03:1. The location of the 118 breakpoints is plotted in Figure 6. It demonstrates that 94 recombination events (80% of the total) could unambiguously be located outside the HLA region, with the HLA complex being defined as the class II, III and I region marked by \(D6S1560\) (300 kb centromeric of \(DPB1\)) on the centromeric and HLA-A on the telomeric side. This is further reinforced by the data listed in Table 2. The overall recombination rate (crossovers/Mb) within the HLA was lower than in the flanking regions by a factor of 0.5–0.6 (compared with the telomeric or centromeric region, respectively). Within the HLA complex, recombination was found clustered between marker \(D6S1560\) and \(DQB1\) (eight recombinant chromosomes in 710 kb; 11.3 breakpoints/Mb). In contrast, we observed only one recombination event within the 850 kb area between \(D3A\) and HLA-C (1.2 breakpoints/Mb) which was located in the 22 kb interval between 82-1 and TNFd. Telomeric of this area recombination frequency increased. Overall, recombination in the class III region was extremely rare in our data set. In the class I region recombination occurred only telomeric of HLA-C.

**DISCUSSION**

We have tested the ability and accuracy of linkage and association analysis in mapping an already identified susceptibility locus of a complex disease, \(IDDM1\), in the HLA class II region, and evaluated the influence of underlying LD patterns, recombination and varying marker informativity. Within the 95% confidence interval obtained by linkage analysis we identified with association analysis a region of 570 kb as the interval most likely to contain the disease gene. Maximal association was found at a marker as close as 85 kb centromeric of the disease-predisposing class II gene \(DQB1\). This sharp peak of association and the correlation of extending association with the LD between markers and \(DQB1–DRB1\) reinforce that the \(DQB1\) and \(DRB1\) loci are the major constituents of \(IDDM1\), dominating the entire HLA complex and its flanking regions.

Multipoint linkage analysis on a detection map with markers physically spaced every 1.3 Mb showed strong linkage of the entire region with LOD scores ranging between 24.8 and 43.3. The broad shape of the curve did not suggest the location of a
linked gene at a glance. Furthermore, calculation of the 95% confidence interval according to Kruglyak and Lander (5) resulted in a ~1.7 cM region which also included the HLA class II region. However, within this 95% confidence interval the linkage curve had a nearly flat shape, implying no subregion of interest within this still fairly large interval.

In contrast, with association analysis a large area of the studied region was unlikely to be a candidate region. After genotyping of 12 additional markers within the previously assessed candidate region (95% confidence interval), the region showing association with significance stronger than \( P = 1 \times 10^{-5} \) comprised an area of ~1.6 Mb. Furthermore, association on both sides of the 570 kb region set by markers D6S2444 and D3A showed a steep drop. Even within this area, a sharp peak of association over marker D6S2444 pointed very precisely to the location of the two major susceptibility genes DQB1 and DRB1. This maximum of association at D6S2444 was further substantiated by the bootstrap procedure, which provided empirical support for the accuracy of the peak of association. The fact that the closest marker to DQB1 and DRB1 among those typed in our study was the most associated makes it likely that a marker with a comparable informativity located closer to these genes would show even stronger association, which is underlined by the extraordinarily strong association of the genes themselves [\( T_c; DQB1 - \log_{10}(P) = 64.7; DRB1 - \log_{10}(P) = 59.0 \)]. Overall, our data show that accurate fine mapping of a major disease locus was possible with polymorphic microsatellites spaced on average every 180 kb despite the extraordinary variation and high level of LD in the HLA region.

No association was found with markers located as close as ~400 kb centromeric of HLA–DQB1. This indicates that even

<table>
<thead>
<tr>
<th>Region Marked by marker/locus</th>
<th>Centromeric of HLA</th>
<th>HLA</th>
<th>Telomeric of HLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>D6S1641–D6S1560</td>
<td>7.1</td>
<td>3.3</td>
<td>3.7</td>
</tr>
<tr>
<td>D6S1560–HLA-A</td>
<td></td>
<td>16</td>
<td>33</td>
</tr>
<tr>
<td>HLA-A–D6S1691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical length (Mb)</td>
<td>Breakpoints/Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of breakpoints</td>
<td></td>
<td>8.6</td>
<td>4.9</td>
</tr>
<tr>
<td>unambiguously within this area</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of recombination rates (breakpoints/Mb) in the HLA and its centromeric and telomeric flanking regions

![Figure 6. The location of 118 recombination events on recombinant chromosomes marked by at least two informative markers on either side of the crossover. Each of the 118 bars represents the area of one recombination on one chromosome. Maternal recombination is indicated by light grey, paternal recombination by dark grey. Orientation of the map is from centromere (left) to telomere (right).](image)
for a susceptibility locus with a large contribution to the familial clustering of the disease, markers may have to be very closely linked to detect its presence when LD is weak. The lack of association and LD observed centromeric of DQB1 corresponds to the presence of three recombination hotspots that have been reported between the DPB1 and DQB1 genes, namely between DNA and RING3, TAP1 and TAP2, and DQB3 and DQB1 (10). With the markers typed in our study these intervals correspond to D6S1560–D6S2445, D6S2445–D6S2444 and D6S2444–DQB1, respectively, although they were mapped on a much finer scale by Cullen et al. (10). Indeed, these intervals were the location where most of the recombination within the HLA region was observed. Conversely, with the exception of one recombination event centromeric of the TNF region, no recombination was found in the HLA class III and the part of the class I region centromeric to HLA-C. This area of low recombination frequency corresponds to the ‘shoulder’ of association (Fig. 2) and strong LD (D’ ≥ 0.4) telomeric of the class II region (Fig. 3). While the drop in disease association telomeric of HLA-C matched increasing recombination rate, the marker to marker LD remained highly significant in this area. These observations underline that recombination patterns in a certain region have to be taken into account for the interpretation of association studies. The recombination female:male ratio of 2:03:1 over the entire 14 Mb studied here corresponds to that of 1:9:1 reported for chromosome 6 (22). The overall recombination frequency within the HLA region was lower compared with its flanking regions on either side (Table 2).

LD in the HLA region has been reported to extend >1 Mb (11). In our study, disequilibrium throughout the HLA region extended in total over an unusual length as far as 5 Mb to the telomeric side. The impressive similarity of the LD analysis of AFBAC chromosomes (data not shown) with the picture shown in Figure 3 emphasizes that this graph represents a good estimate of general population LD in the HLA and its flanking region. Thus, the data presented here can be of great importance for the mapping of other HLA-associated disease loci. It should be noted that the region telomeric of the HLA complex showed significant extended LD whereas the region centromeric of the HLA did not (Fig. 3), although both these areas showed a similar density of recombination events (Fig. 6). A possible explanation for this phenomenon could be selection processes operating between the HLA class I region and the telomeric end, leading to a reduction in haplotype diversity (11). Furthermore, our more detailed analyses demonstrate how the association curve follows the LD between markers and the extended DQB1–DRB1 susceptible haplotypes (Figs 4 and 5). Therefore, additional peaks of association outside the class II region are likely to be caused by LD with DQB1–DRB1. Obviously, prior to a fine mapping study, this information is not available, but as shown here, the LD pattern between markers can predict the association observed. This shows that prior information about LD in a certain region can be helpful for designing a mapping experiment and is important for the correct interpretation of association analysis. With the progress of the human genome project valuable and necessary data on LD, extended haplotypes and recombination patterns of any region in the human genome will become available in electronic databases.

One may argue that it would be hard to not detect a major disease locus like IDDM1 in a region with such strong LD. In fact, any of the markers studied here in the class II and III regions would have picked up a strong association signal when used in the detection map. The sibling/population prevalence risk ratio λs for HLA in type 1 diabetes was 2.7 in this UK data set. This is a single large contribution to the overall λs of 15 but other HLA-associated diseases show similarly strong or even stronger HLA association, e.g. ulcerative colitis or rheumatoid arthritis (23). Beyond autoimmune disease, it is anticipated that loci with λs values in the range 2–3 will be found, and published power calculations often include this range (24). Moreover, the principles and basic properties of association mapping discussed here will apply to disease loci with smaller λs values, except that much larger data sets will be required. However, the focus of this study was not the detection but rather the accuracy of mapping the IDDM1 locus, which was possible with a surprisingly high resolution despite the strong extended LD. Two recent case–control studies have narrowed down HLA candidate regions to intervals of 46 kb for Behçet disease and 111 kb for psoriasis using microsatellite markers (25,26). This shows that the mapping of disease loci weaker than IDDM1 also seems possible within the HLA region. However, the aetiological genes in these studies have not yet been determined and the investigation of LD with the disease locus was consequently not possible. Furthermore, no attempt was made in those studies to define a confidence interval for the association.

The CETDT (20) allows us to test for association while correcting for LD with another associated locus or haplotype. This test confirmed that the highly significant association observed in the HLA class III region was not due to an independent effect but was caused by LD with the DQB1 and DRB1 loci. However, D6S2223, located 5.9 Mb telomeric of DRB1, showed independent association on the DR3 susceptible haplotype after correction for LD. Lie et al. (21) found allele 3 of D6S2223 on DR3 haplotypes to be associated with reduced susceptibility in families with type 1 diabetes using the homozygous parent TDT. They reproduced this finding for coeliac disease in a case control study (27). Although in our data set the homozygous parent TDT showed a low transmission rate of allele 3 (181 bp) at D6S2223 on the DQB1*02–DRB1*03 haplotype, numbers were too low to show a significant result. This confirms that very large data sets are necessary for this test. However, our conditional ETDT results emphasize evidence for a secondary gene in LD with D6S2223 modulating diabetes susceptibility independently from the exon 2 sequence of DQB1 and DRB1, although these data cannot be seen as a replication of Lie’s results since the families investigated here were overlapping with those analysed by Lie et al. (21).

Apart from varying LD, markers differ in the number of parents that are informative (i.e. heterozygous). A possible way to balance informativity between markers is to use two-point marker haplotypes, which may increase heterozygosity and reduce, to a certain extent, the variation in strength of association owing to marker informativity. As we showed, association was stronger for the D6S2444–D3A haplotype than for the single markers correlating with an increase in heterozygosity. The two-point haplotypes also increased association of markers 9N-1 and D6S273, located near the TNF gene. For
these markers, however, two-point haplotypes did not substantially improve heterozygosity. Instead, the LD between the two-point haplotypes and the susceptible DQ–DR haplotypes was stronger than that for the single markers. These observations confirm that the use of haplotypes may in some instances improve association mapping by increasing informativity and/or LD between marker and disease locus as suggested previously (28).

Our study demonstrates that microsatellite markers can be a valuable tool for the fine mapping of a multifactorial disease locus. However, very large data sets may be required for the mapping of weaker loci and consequently polymorphism scoring techniques that are subject to automation are necessary. Since microsatellite markers are not suitable for fully automated scoring, single nucleotide polymorphisms (SNPs) are anticipated as the marker of choice (29). However, the utility of microsatellites versus SNPs for LD mapping of susceptibility awaits the analysis of data sets with both microsatellites and SNPs typed.

Overall, an association study can benefit greatly from prior knowledge about the LD pattern of the region under study. This information may be used to choose marker density appropriately, and is necessary for interpretation of the association obtained together with the informativity index of markers. With this information association can lead to accurate mapping of complex disease loci. Although there has been concern that the location of a disease-predisposing locus may not be located with much accuracy (6,9), we demonstrate in real data that even in a chromosome region with extraordinary levels of variation in LD the disease locus may indeed be mapped very accurately.

MATERIALS AND METHODS

Diabetic families and HLA typing

The 385 families used in this study were all Caucasian Europeans with grandparents born in the UK, and consisted of affected sib-pair families with one sibling diagnosed with type 1 diabetes before the age of 17 years and the other sibling before the age of 29 years, as described (30). Family samples were obtained, together with their HLA-typing, as part of the British Diabetic Association–Warren Repository (30).

Genotyping microsatellite markers

All families, parents and offspring, were genotyped for 25 microsatellite markers. The primer sequences for D6S1641, D6S1548, D6S1576, D6S291, D6S439, D6S1629, D6S1560, D6S1568, D6S2445, D6S2444, D6S273, C1-2-A, C1-2-C, C1-4-4, C1-3-2, D6S265, D6S258, D6S1683, D6S306 and D6S1691 were obtained from Foissac and Cambon-Thomsen (31). Sequences for D3A, 82-1, 82-2/9N-1 and 62b were established in the laboratory of D. Campbell (32). Primers for 82-2 and 9N-1 were found to amplify the identical repeat sequence, while 10 parents were grouped together for the analysis with Tsp, ETDT. Alleles and haplotypes occurring in fewer than 73% of all parents were heterozygous for the microsatellites investigated, with outlying markers D6S1576, D6S2445 and D6S2223 showing low heterozygosity <60%. The average number of alleles per microsatellite was 12, but only three for alleles with a parental frequency of at least 10%.

Physical map of microsatellites

The physical microsatellite map was established by locating the primer sequences on the consensus sequence of the HLA region released by the Sanger Centre (http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml ) and for the flanking regions with the help of the Sanger Centre and information from Foissac and Cambon-Thomsen (31) and Feder et al. (33).

Linkage, association and recombination analysis

Multipoint linkage analysis was performed as a maximum likelihood estimate of IBD sharing using the ‘estimate’ function of the MAPMAKER/SIBS 2.0 software (34). The analysis was based on a genetic multipoint distance map, calculated with the sibmap program of the ASPEX software package (D. Hinds and N. Risch, available at ftp://lahmed.stanford.edu/pub/aspx ).

Association was tested with the two TDT-based multi-allelic statistics Tsp (15) and ETDT (16), which correct for linkage within sib-pair families. Tsp is a test of marginal heterogeneity with a correction of non-independence of affected sib pairs; ETDT is a logistic-regression formulation of the TDT for multiple alleles, to which we have applied the Tsp correction (15) for linkage in sib-pair families, which only counts transmissions of alleles transmitted to both siblings. The CETDT (20), which is an adaptation of the ETDT, was used to test association corrected for LD with another locus or haplotype, in this study the DQB1–DRB1 haplotype. The CETDT is based on the principle that haplotypes with identical alleles at a candidate disease locus (or most associated marker), but different alleles at the test locus, have equal transmission probabilities. The ETDT method can be used to estimate transmission probabilities of such haplotypes from family data, and are tested for equality by CETDT. Alleles and haplotypes occurring in fewer than 10 parents were grouped together for the analysis with Tsp, ETDT and CETDT. When haplotypes were unknown due to phase ambiguity, the data was discarded with allowance made for possible bias (35). The association tests were performed using computer programs written by F.D. Programs are available by anonymous ftp (http://diesel.cimr.cam.ac.uk ) or by contacting F.D. (frank@gene.cimr.cam.ac.uk ). The majority of P-values are given (Perkin-Elmer Applied Biosystems, Warrington, UK). The alleles at each microsatellite were given a numerical value (1, 2, 3, etc.) starting with the allele with the lowest number of base pairs. PCR product standards, consisting of the amplification product of two different standard individuals for each marker, were loaded on each gel for correct allele assignment. Characteristics and physical map position of the microsatellite markers are shown in Table 1. Markers were PCR amplified and genotyped a second time when showing misinheritance within a family. Furthermore, families were checked for multiple recombination. Families showing persistent misinheritance or tight double recombination, indicative of possible genotyping errors or mutation, after a second round of genotyping, were discarded from analysis in this study. On average 73% of all parents were heterozygous for the microsatellites investigated, with outlying markers D6S1576, D6S2445 and D6S2223 showing low heterozygosity <60%. The average number of alleles per microsatellite was 12, but only three for alleles with a parental frequency of at least 10%.
as their negative log10 in the text, designated $-\log_{10}(P)$. Recombination was analysed using the sibmap function of the GAS software package version 2.0 © Alan Young, Oxford University, 1993–1995. This assesses crossovers by changes in allele sharing status between siblings given a certain marker order. For the recombination analysis we additionally used genotypes for the microsatellites TNF$a$, TNF$b$, and TNF$c$ [primer sequences obtained from Udalova et al. (36)] and the HLA class I loci B, C and A.

**LD analysis**

LD was measured using two different statistics: pairwise D′ and total D′. The pairwise D′ and significance were calculated using the ARLEQUIN software package (available at http://anthropologie.unige.ch/arlequin ). Pairwise D′, an allele-specific measure, ranges from –1 to +1, with –1 and +1 representing complete LD and 0 representing free association. The total normalized disequilibrium (total D′), a multiallelic measure, was calculated according to equation 14 in ref. (37). This measure ranges from 0 to 1, with 0 reflecting no and 1 reflecting complete LD. Parental gametic haplotypes were deduced from parental and offspring genotypes using software written by F.D. We used both transmitted and non-transmitted haplotypes to one sibling per family in these calculations. As a control, total LD was also investigated in AFBAC chromosomes (19).

**ACKNOWLEDGEMENTS**

We would like to thank Andrew Mungall (Sanger Centre) for his help with establishing the microsatellite map, and Iain Eaves for important help with the LD analysis and for making a computer program available for the calculation of total D′. This research was supported by the Wellcome Trust, the Juvenile Diabetes Foundation and by Eli Lilly Industries, UK. M.H. was a postdoctoral research fellow funded by the Juvenile Diabetes Foundation International (JDFI). B.P.C.K. was funded by a TMR Marie Curie Research Training Grant of the Diabetes Foundation International (JDFI). B.P.C.K. was funded by a Wellcome Trust Principal Grant and the Medical Research Council. J.A.T. was a Wellcome Trust Principal funded by a TMR Marie Curie Research Training Grant of the European Commission. J.A.T. was a Wellcome Trust Principal funded by a TMR Marie Curie Research Training Grant of the Diabetes Foundation International (JDFI). B.P.C.K. was a computer program available for the calculation of total D. A m . J. H u m . G e n e t . 513–519.

**REFERENCES**
