Haplotype tagging for the identification of common disease genes

Gillian C.L. Johnson[1], Laura Esposito[1], Bryan J. Barratt[1], Annabel N. Smith[1], Joanne Heward[2], Gianfranco Di Genova[1], Hironori Ueda[1], Heather J. Cordell[1], Iain A. Eaves[1], Frank Dudbridge[1], Rebecca C.J. Twells[1], Felicity Payne[1], Wil Hughes[1], Sarah Nutland[1], Helen Stevens[1], Phillipa Carr[1], Eva Tuomilehto-Wolf[3], Jaakko Tuomilehto[3,4], Stephen C.L. Gough[2], David G. Clayton[1] & John A. Todd[1]

Genome-wide linkage disequilibrium (LD) mapping of common disease genes could be more powerful than linkage analysis if the appropriate density of polymorphic markers were known and if the genotyping effort and cost of producing such an LD map could be reduced. Although different metrics that measure the extent of LD have been evaluated[1–3], even the most recent studies[2,4] have not placed significant emphasis on the most informative and cost-effective method of LD mapping—that based on haplotypes. We have scanned 135 kb of DNA from nine genes, genotyped 122 single-nucleotide polymorphisms (SNPs; approximately 184,000 genotypes) and determined the common haplotypes in a minimum of 384 European individuals for each gene. Here we show how knowledge of the common haplotypes and the SNPs that tag them can be used to (i) explain the often complex patterns of LD between adjacent markers, (ii) reduce genotyping significantly (in this case from 122 to 34 SNPs), (iii) scan the common variation of a gene sensitively and comprehensively and (iv) provide key fine-mapping data within regions of strong LD. Our results also indicate that, at least for the genes studied here, the current version of dbSNP would have been of limited utility for LD mapping because many common haplotypes could not be defined. A directed re-sequencing effort of the approximately 10% of the genome in or near genes in the major ethnic groups would aid the systematic evaluation of the common variant model of common disease.

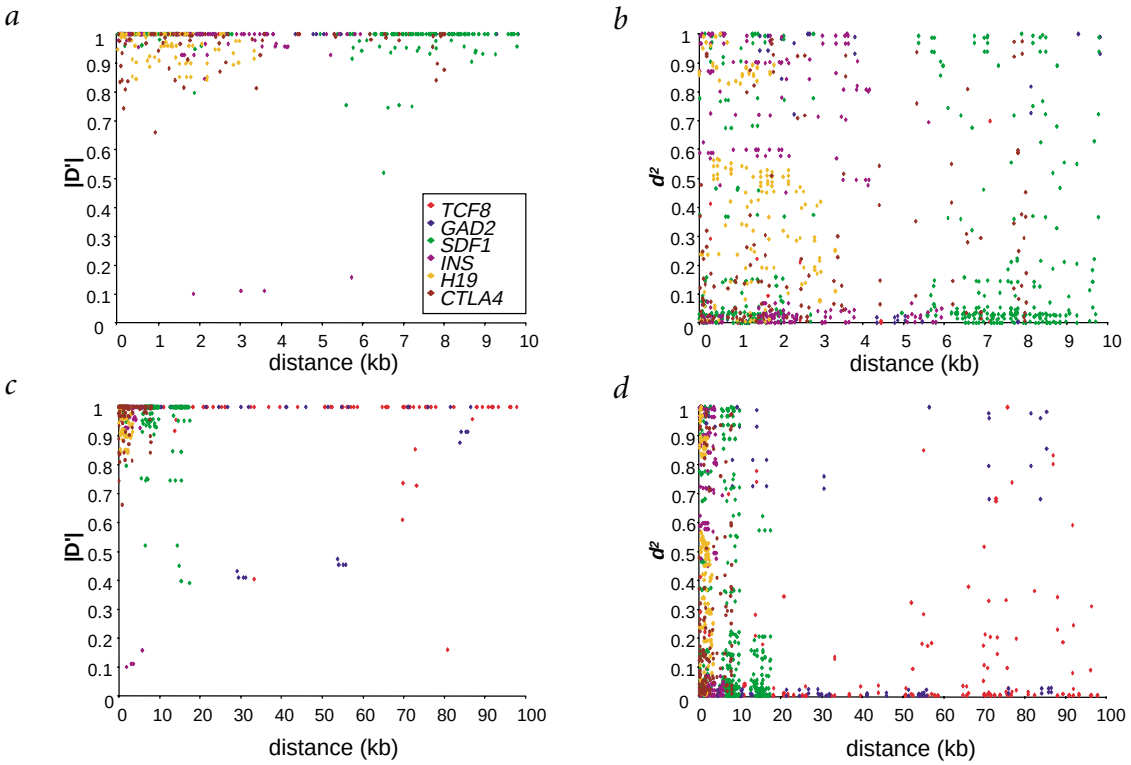

Fig. 1. Relationship between |D'|, d^2 and physical distance. ***a,c***, |D'| values for all pairs of markers genotyped in *TCF8, GAD2, CTLA4, SDF1, INS* and *H19* related to the distance between the two loci (10-kb window and100-kb window). ***b,d***, Pairwise d^2 values (10-kb window and 100-kb window).

[1]*JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/Medical Research Council Building, Hills Road, Cambridge, UK.* [2]*Department of Medicine, University of Birmingham and Birmingham Heartlands and Queen Elizabeth Hospitals, Birmingham, UK.* [3]*Diabetes and Genetic Epidemiology Unit, National Public Health Institute, Helsinki, Finland.* [4]*Department of Public Health, University of Helsinki, Mannerheimintie, Helsinki, Finland. Correspondence should be addressed to J.A.T. (e-mail: john.todd@cimr.cam.ac.uk).*

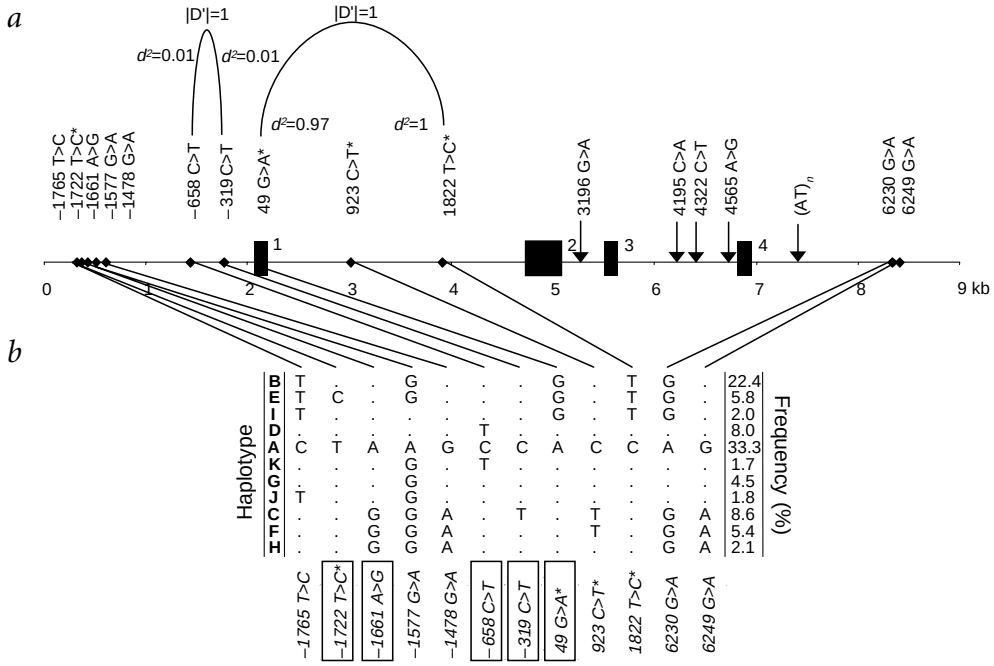

Fig. 2 Polymorphisms detected and genotyped at *CTLA4*. ***a***, Coding exons are marked by shaded blocks. Genotyped polymorphisms are marked as a diamond on the diagram; the positions of untyped polymorphisms (including the $(AT)_n$ microsatellite and the four SNPs that were heterozygous in only one individual in our screening panel (n=46)) are marked by arrows. Polymorphisms marked with an asterisk were also described in dbSNP. Base A of the ATG of the initiator Met codon of *CTLA4* is denoted nucleotide +1. ***b***, Dots represent the allele that is found on the most common haplotype. Five haplotype-tagging SNPs (htSNPs; boxed) describe all of the common (>5% frequency) haplotypes we observed in the region.

Intense effort has been put into determining the best metric to measure LD[1–3]. Figs. 1 and 2 show the lack of correlation between the level of LD and physical distance in regions smaller than 100 kb when using two common measures, D′ and d^2 (Methods). If, however, the underlying haplotypes are characterized, either by reconstructing parental haplotypes from family pedigree data or by estimating haplotype frequencies from the genotypes observed in unrelated individuals, the relationships between all alleles in the region can be clearly defined (Fig. 2). For example, at the *CTLA4* locus, allele T of marker –319C→T (–319C→T*T) occurs only with allele C of marker –658C→T (on haplotype C), as indicated by a D′ value of 1 between these two markers (confirmed by assessing LD between these two markers in 295 European families; data not shown). The d^2 value, however, is very low, because –658C→T*C is also present on several other common haplotypes (A–C and E–J). Consequently, in an association study, marker –319C→T would fail to detect association with disease if marker –658C→T were the disease locus and were not genotyped. These two markers are less than 400 bp apart. In contrast, allele 49G→A*G is found exclusively on the same haplotypes as 1822T→C*T; as a result, both D′ and d^2 values approach 1 between these two markers. One of these markers is, therefore, effectively redundant in a first-pass disease association study.

By determining the extended haplotypes at any given locus in a population, we can identify exactly which SNPs will be redundant and which will be essential in association studies. We refer to the latter as 'haplotype tag SNPs (htSNPs)', markers that capture the haplotypes of a gene or a region of LD (Fig. 3). Although htSNPs can be identified by eye (Figs. 2,3), we have also developed a program that identifies the best groups of htSNPs that capture the majority of the haplotype diversity observed within a region. For the nine genes studied here, 2–5 htSNPs can be used to define the six or fewer common haplotypes (greater than 5% population frequency) observed at each locus. These common haplotypes and their htSNPs account for at least 80% of all haplotypes that we observed. Similarly, it is reported elsewhere in this issue that a small number of common haplotypes account for the majority of Canadian chromosomes observed within regions of the cytokine gene cluster on chromosome 5q31 (refs. 5,6). Although other studies of intra-European haplotype diversity are based on small samples, they indicate that the most frequent haplotypes are shared between the general European-derived populations, as expected[7–11]. In the present study, the four most frequent haplotypes observed when typing six SNPs from the *GAD2* gene are common to both the Finnish and UK populations (with similar frequencies; data not shown). Common haplotypes from different general populations in Europe today will, therefore, be characterized by the same htSNPs. We have identified the htSNPs in five additional genes for which haplotypes of European origin were characterized previously (*ATM*[10], *LCT*[11], *ACE*[12], *MC1R*[13] and *APOE*[14]). A maximum of five htSNPs could be used to capture all of the common variation at each of these loci. Even the *LPL* gene, previously reported to have high levels of sequence and haplotype diversity[9,15], can be tagged with 4–6 SNPs, provided the 'hot spot' of recombination in the gene is taken into account[16,17]. Predetermination of haplotypes and subsequent selection of htSNPs thus promises to significantly reduce the genotyping effort when evaluating the association of common variants with common disease. More importantly, genotyping of the htSNPs ensures that all of the common variation within a region of LD is surveyed in a disease association study. This cannot be guaranteed when an association map is based on SNP spacing alone (such as one SNP per 10 kb) or when adjacent SNPs are combined in a two-locus LD analysis without determining the underlying haplotype structure of a region at the outset[18].

In the analysis of a positional and/or functional candidate gene or region, the optimal experimental design would be to contiguously screen the entire genomic sequence to detect all the common variants that exist. This would ensure that all the common haplotypes are defined. However, given that re-sequencing is time-consuming and expensive, the search for disease-associated SNPs and haplotypes could, in the first instance, be restricted to DNA-containing exons and the immediate 5′ and 3′ region of genes. Indeed, excluding the genes reported here (*CTLA4*, *INS*[19]and *H19*) and in the 5q31 region reported in this issue[5,6], only *ACE*, spanning 24 kb, *APOE*, spanning 5.5 kb, *NOD2/CARD15*, spanning approximately 36 kb, and *CAPN10*, spanning 66 kb, have been completely re-sequenced in several individuals[12,20–22]. We selected the SNPs that

either map within 500-bp fragments containing coding exons or within up to 3 kb 5′ of the ATG of exon 1 and 3 kb downstream of the 3′ untranslated region (UTR) at *CTLA4* (8.6 kb), *INS* (6.5 kb), *H19* (4.4 kb), *APOE*[20] and *ACE*[12]. In all cases, restricting haplotype analysis to those SNPs detected in the targeted regions leads to the detection of suitable htSNPs and, therefore, a complete analysis of the common haplotypes with disease. It is still unclear, however, how restricting polymorphism detection in this way will affect the haplotype characterization of larger genes (>100 kb), as the screened sequences will account for a much smaller proportion of the total genomic sequence of the gene. In addition, some regions of the genome (such as the HLA region) will probably show higher levels of haplotype diversity in Europeans.

We searched dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) to evaluate how many of the SNPs detected here were present in the database (we excluded *INS* from this analysis, as a comprehensive polymorphism screen of the gene has already been published[19]). dbSNP contained no more than 25% of the SNPs that we identified at any one gene (Fig. 3). At the *TCF8*, *H19*, and *CASP10* genes, dbSNP did not contain any of the variants that we observed. For the remaining five genes, *CFLAR*, *CASP8*, *GAD2*, *CTLA4* and *SDF1*, the subset of our SNPs that were described in dbSNP could not distinguish between the common haplotypes that we observed. Although some additional variants were described in the database, these may represent either very rare variants that our screening panel ($n \leq 46$) did not have the power to detect, or sequences that are not polymorphic in the populations we studied[23].

For an initial gene-based catalog of htSNPs, we suggest that instead of relying on dbSNP, at least all of the coding and 3 kb of up- and downstream sequences of each gene should be re-sequenced in a minimum of 30 individuals. This would result in a greater than 95% power to detect all variants with frequencies higher than 5%. This gene-based SNP harvesting approach is already being taken for Japanese chromosomes (http://snp.ims.u-tokyo.ac.jp). A parallel effort for other ethnic groups should be a priority, followed by genotyping of a core panel of DNA samples, enabling the compilation of a genome-wide set of htSNPs for haplotype-based disease association mapping. An htSNP map will allow effective evaluation of the common disease–common variant hypothesis.

For common haplotypes (>5% frequency), attainable sample sizes provide sufficient statistical power to detect susceptibility variants with odds ratios of 1.5 or greater. We chose 5% as the threshold between common and rare haplotypes because sample-size requirements increase dramatically when allele frequencies fall below 5%. For example, assuming a multiplicative model, an equal number of either 5,893 or 27,508 cases and controls would be required to have 80% power to detect a disease variant with an odds ratio of 1.5 at $P \leq 10^{-5}$ if the disease allele had frequencies of 5% or 1%, respectively (Fig. 4). If suitably powered studies that evaluate common variation fail to produce convincing associations, the multiple-rare-variants model of common disease will then need to be considered. For example, in the recently discovered Crohn disease susceptibility gene (*NOD2/CARD15*), a number of rare variants clustered within the gene underlie disease susceptibility[22,24]. Statistical tests that combine information from all of the rare variants of a gene may facilitate the detection a disease locus comprised of several rare alleles (D.G.C. & J.A.T., unpublished data)[25].

CFLAR

C/T	A/T	A/G*	T/–	G/T	G/A	Freq
C	A	A	T	G	G	46%
.	.	G	.	T	A	44.25%
T	.	G	.	T	A	8.75%
.	T	G	-	.	A	0.5%

CASP10

C/T	C/T	A/G	A/G	C/T	G/A	A/G	T/A	G/A	G/C	G/A	Freq
C	C	A	A	C	G	A	T	G	G	G	44%
T	.	G	.	.	.	A	.	C	A		39%
T	.	G	.	.	A	.	A	.	C	A	7%
T	.	G	.	.	.	G	A	.	C	A	6.25%
.	T	G	G	T	.	.	.	A	.	A	1.75%
.	T	G	G	.	.	.	A	.	C	A	0.5%

GAD2

A/G	C/T	C/A	A/G	A/G	C/A	G/A	A/G	C/A/T	T/G	C/T	G/C*T/A*	Freq
A	C	C	A	G	C	G	A	C	T	C	G	T	45%
.	.	.	.	.	.	.	.	T	.	.	.	.	28%
.	T	A	G	A	.	.	.	T	.	.	.	A	12.75%
G	.	.	.	.	.	.	G	A	G	T	C	.	8.25%
.	.	.	.	.	.	.	.	T	.	.	.	A	1.75%
.	.	.	.	.	.	A	.	T	.	.	.	.	1%
.	T	A	G	A	.	.	.	T	.	T	C	.	0.75%
.	.	.	.	.	A	.	.	T	.	.	.	.	0.50%

H19

G/C	G/C	G/T	T/C	C/T	C/T	A/T	A/G	C/G	G/C	G/T	G/A	G/A	Freq
G	G	G	T	C	C	A	A	C	G	G	G	G	34.75%
C	.	T	C	T	.	T	G	G	C	T	A	A	19%
.	C	.	C	T	.	T	G	G	C	T	A	A	15%
.	.	.	.	.	.	.	.	.	.	.	A	A	10.25%
.	.	T	C	T	.	T	G	G	C	T	A	A	5.75%
.	.	.	.	.	T	.	.	.	.	.	.	.	4.50%
.	C	.	C	T	.	T	G	.	C	T	A	A	1%
.	.	.	C	T	.	T	G	G	C	T	A	A	1%
.	.	.	.	.	.	.	.	G	C	T	A	A	1%
C	.	T	C	T	.	T	G	G	C	T	A	.	0.5%
.	C	.	C	T	.	T	G	.	C	T	A	.	0.5%
.	C	.	C	T	.	T	G	G	C	T	.	A	0.5%
.	.	.	C	T	.	T	G	G	C	T	A	.	0.5%
.	.	.	.	.	.	.	.	.	C	.	.	.	0.5%
.	.	.	.	.	.	.	.	G	C	.	A	A	0.5%

INS

A/C	C/T	A/T	C/G	C/T	C/T	C/A	C/T	G/T	G/A	G/A	C/A	C/T	G/A	Freq
A	C	A	C	C	C	C	C	G	G	G	C	C	G	45%
.	.	.	.	.	.	.	.	.	.	.	A	.	.	20%
C	T	T	G	.	T	A	T	T	A	.	.	.	.	13.25%
.	.	.	.	.	.	.	.	.	.	.	.	.	A	11.25%
C	.	T	.	T	.	A	.	.	.	A	A	.	.	3.75%
.	.	.	.	.	.	.	.	.	.	.	.	T	.	3.50%
C	.	.	.	.	.	.	.	.	.	.	.	.	.	1.50%
C	.	T	.	T	.	A	.	.	.	.	.	.	.	0.5%
.	T	T	G	.	T	A	T	T	A	.	.	.	.	0.5%

TCF8

G/A	A/G	C/T	T/C	T/C	T/G	C/G	T/C	A/G	T/C	A/G	G/A	T/C	A/G	Freq
G	A	C	T	T	T	C	T	A	T	A	G	T	A	33.5%
.	.	T	.	C	.	.	.	.	.	G	.	.	.	13.75%
.	.	T	C	C	.	.	.	.	.	.	.	.	.	13.25%
.	.	T	C	C	.	.	C	.	.	.	.	C	.	8.25%
.	.	T	.	C	.	.	.	.	.	.	.	.	.	8%
.	.	T	C	C	.	.	.	.	.	.	.	C	.	5.25%
.	.	.	.	C	.	G	.	.	.	.	A	.	.	4.5%
A	.	.	.	.	.	.	.	.	.	.	.	.	.	3.75%
.	G	T	.	C	.	.	.	.	.	G	.	.	.	2.25%
.	.	.	.	.	.	.	.	G	.	.	.	.	.	1.75%
.	.	T	C	C	G	.	.	.	.	.	.	C	.	1.75%
.	G	T	T	C	.	.	.	.	C	G	.	.	.	1.25%
.	.	.	.	.	.	.	.	.	.	.	.	.	G	0.75%
.	.	.	.	C	.	.	.	.	.	.	.	.	.	0.75%
.	.	T	C	.	.	.	.	.	.	.	.	.	.	0.75%

CASP8

T/G	T/C	G/A	G/T	C/G*	G/A	G/A	C/G	C/T	G/C	A/G	A/G	C/A	Freq
T	T	G	G	C	G	G	C	C	G	A	A	C	39%
.	.	A	.	.	A	A	G	.	.	.	.	A	18.25%
.	.	.	.	G	.	.	.	.	.	.	G	.	12.75%
.	C	.	.	.	.	.	.	.	.	.	.	.	9.25%
.	.	A	.	.	A	.	G	.	.	.	.	A	6.50%
.	.	A	.	.	A	.	G	.	.	G	.	A	3.75%
.	.	.	.	.	.	.	.	.	.	.	.	A	2.75%
G	.	A	T	.	.	.	.	.	.	.	.	.	1.75%
.	.	.	.	.	.	.	G	T	C	.	.	A	1.75%
.	.	.	.	.	.	.	G	T	.	.	.	A	1.25%
.	.	.	.	.	A	.	G	.	.	.	.	A	1%
G	.	A	T	.	.	.	G	.	.	.	.	A	0.75%

SDF1

G/A	A/G	C/T	G/C	G/A	G/A	C/T	C/T	A/G	G/C	A/T	T/C	+/–	T/C	G/A	+/–*	T/C*	G/A*	C/T*	T/C*	C/T	C/G*	Freq
G	A	C	G	G	G	C	C	A	G	A	T	+	T	G	+	T	G	C	T	C	C	30.75%
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	A	T	.	.	.	17.50%
.	.	.	.	.	.	T	.	.	C	.	.	.	.	A	-	.	.	T	.	.	G	15%
A	G	.	.	A	.	.	T	G	.	T	C	.	C	.	-	C	.	T	.	.	G	14.25%
.	G	.	C	.	A	.	T	.	.	.	.	.	.	.	-	.	.	T	.	.	.	10%
.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	T	.	.	.	5.5%
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	T	.	.	.	1.5%
.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	-	.	.	.	C	T	G	1.5%
A	G	.	.	A	.	.	T	G	.	T	C	.	.	.	-	C	.	T	.	.	G	0.75%
A	G	.	.	A	.	.	T	G	.	T	C	.	C	.	-	C	A	T	.	.	.	0.5%

Fig. 3 Common European haplotypes and their htSNPs observed at nine genes. Boxed SNPs represent the htSNPs that can capture the common haplotypes that are segregating in European populations. Dots represent the allele that is found on the most common haplotype. Asterisks indicate SNPs described in dbSNP.

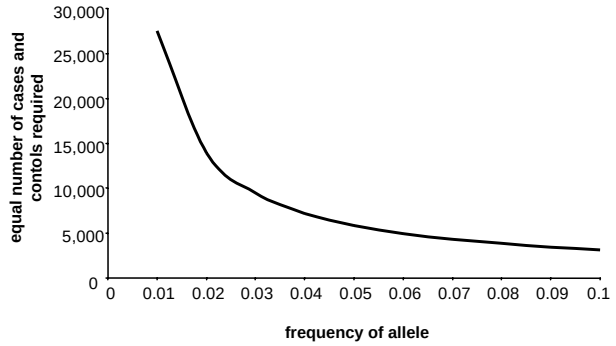

Fig. 4 Relationship between allele/haplotype frequency and sample-size requirements for population-based association studies. We calculated the sample sizes for a population-based study assuming a multiplicative model and based on having 80% power to detect a disease effect of odds ratio 1.5 at a significance level of $P=1\times10^{-5}$. These calculations assume that the etiological variant can be defined by the htSNPs. The numbers of trio families (both parents and the proband) required are almost identical, and, therefore, require one-third more genotyping effort.

Studying haplotypes has two further advantages. If the disease association of a specific allele is dependent on *cis* interactions with other loci, the disease association may not be detected unless the functional haplotypic unit itself is analyzed[21,26–28]. In addition, exploiting differences in haplotype diversity and frequency between populations ('trans-racial mapping'[29]) may be invaluable when attempting to pinpoint which variants are most likely to be the primary etiological determinants of common diseases[4,29–31].

Methods

Subjects. We genotyped polymorphisms in *INS*, *H19*, *SDF1*, *TCF8* and *GAD2* in a maximum of 418 multiplex families originating from the UK in which at least two siblings were diagnosed with type 1 diabetes (the Diabetes UK Warren 1 repository). We constructed haplotypes at *CASP8*, *CASP10* and *CFLAR* using 598 Finnish families in which at least one sibling had been diagnosed with type 1 diabetes (L.E. *et al.*, unpublished data). We genotyped markers at *CTLA4* in 384 unrelated white UK individuals with no history of autoimmune disease who gave blood at various sites, including the Blood Transfusion Service, Birmingham Heartlands Hospital and the Queen Elizabeth Hospital, Birmingham.

Polymorphism detection. We screened at least 20 parents of type 1 diabetic sibs from the UK and at least 10 white UK control individuals who were not ascertained for disease status for polymorphisms in the *TCF8*, *GAD2*, *SDF1* and *INS* genes using D-HPLC (Transgenomic WAVE). We detected variants in *H19* by screening at least 20 parents of type 1 diabetic sibs from the UK, again by D-HPLC. We characterized polymorphisms at *CTLA4* using a screening panel comprised of 10 parents of type 1 diabetic sibs from the UK, 10 parents from diabetic simplex families from both Finland and Italy and 16 white control individuals from the UK who were not ascertained for disease status (S.C.L.G. & J.A.T., unpublished data). We detected variants in *CASP8*, *CASP10* and *CFLAR* by re-sequencing 10 parents from diabetic simplex families from Finland and 20 parents of type 1 diabetic sibs from the UK (L.E. *et al.*, unpublished). All regions screened were within 5 kb of the 5′ untranslated region (UTR) and within 3 kb of the 3′ UTR of each gene.

Genotyping. We genotyped the majority of polymorphisms using Invader[32], by conventional restriction fragment length polymorphism (RFLP) assays or by amplification-refractory mutation system (ARMS)–PCR.

Statistical analysis. We reconstructed parental haplotypes from pedigree data using software available from our website (ftp://ftp-gene.cimr.cam.ac.uk/software). For each family, we determined the parental origin of the offspring alleles for all possible loci. Where appropriate, we used the genotype information of a tightly linked marker to resolve phase. We treated cases in which phase could not be unambiguously resolved as if the genotype information were missing. By choosing one offspring from each family at random, we could identify four unrelated chromosomes, which, under the assumption of no recombination, represented the parental chromosomes themselves. For 8 of the genes studied, we then selected 400 fully genotyped haplotypes at random from the available data. We omitted haplotypes that were observed only once from further analysis to avoid potential genotyping error. We genotyped markers at *CTLA4* in 384 unrelated individuals and estimated haplotype frequencies using the expectation-maximization (EM) algorithm. Any haplotype that had a frequency of less than 1% was excluded from further analysis to avoid possible errors in either the genotyping or the estimation process. We quantified LD between all pairs of biallelic loci using the absolute unsigned value of Lewontin's D' (|D'|) and the measure d^2. Both measures can take on values from 0 (no LD) to 1 (complete LD). However, the scale of |D'| is independent of allele frequency such that two alleles in complete LD can exhibit |D'|=1 when their frequencies are divergent; that is, when only three of the four possible haplotypes are observed. In addition, d^2=1 is only attained when two alleles are in absolute LD and have identical frequencies; that is, when only two of the four possible haplotypes are observed. We calculated d^2 as $D^2/[f(1-f)]^2$, where f is the frequency of the putative disease variant. As either locus in each pair could be assumed to represent the putative disease locus, f could take either the frequency of allele i at locus one or allele j at locus two (all markers tested were biallelic). Therefore, we calculated both possible values of d^2. To assist in the selection of htSNPs, we initially ordered haplotypes in terms of their similarity using ClustalX software (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/) and selected the htSNPs by eye. A more formal approach is to search all possible subsets of markers to identify that subset of a given size that best captures the full haplotype information. What comprises the 'best' choice is open to debate. We have defined optimality in terms of the residual haplotype diversity within groups of haplotypes identified by the htSNPs. Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes, and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus. We have investigated choice of htSNPs so as (i) to maximize the overall percent of diversity explained and (ii) to maximize the worst of the locus-specific values. In all cases, there was a fairly wide choice of htSNP selections with very similar performance, and we had no difficulty selecting sets that performed well in both respects. Full details of the procedure may be found on our web site (http://www-gene.cimr.cam.ac.uk/clayton/software/stata). The ability to predict a further SNP from a group of htSNPs is measured by the locus-specific index of 'diversity explained', which is calculated by our program. Using a small number of htSNPs, we had no difficulty achieving high levels of this index for the great majority of common loci we studied.

Power calculations. We calculated the required sample size to achieve a given power as follows. Suppose we have n case chromosomes and n control chromosomes. Let the allele frequency of the associated allele be q and p in case and control chromosomes respectively. Under the null hypothesis of no effect, this allele frequency can be estimated by

$$r=\frac{(p+q)}{2}$$

the log odds ratio (ln OR) is assumed to be 0 and the estimated standard error of ln OR is

$$\frac{\sqrt{\left[\frac{2}{r}+\frac{2}{(1-r)}\right]}}{\sqrt{n}}=\frac{S_0}{\sqrt{n}}$$

Under the alternative (true) hypothesis, ln OR is

$$\frac{q(1-p)}{p(1-q)}=\mu_1$$

with standard error

$$\frac{\sqrt{\left[\frac{1}{q}+\frac{1}{p}+\frac{1}{(1-q)}+\frac{1}{(1-p)}\right]}}{\sqrt{n}}=\frac{S_1}{\sqrt{n}}$$

Given p and the OR, we may therefore calculate q, r, μ_1, S_0 and S_1, and n as

$$n=\frac{[Z_{1-\alpha}S_0+Z_{1-\beta}S_1]^2}{\mu_1^2}$$

where $Z_{1-\alpha}$ and $Z_{1-\beta}$ are the critical values for the standard normal distribution corresponding to significance level α and power $1-\beta$.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

We are grateful to L. Smink for advice. This work was funded by the Wellcome Trust, the Juvenile Diabetes Research Foundation International, Diabetes UK and grants from the Finnish Academy (38387, 46558, 52114, and 51225), Novo Nordisk Foundation and Sigrid Juselius Foundation. G.C.L.J. was the recipient of a Diabetes UK PhD Studentship.

Received 28 June; accepted 24 August 2001.

1. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
2. Morton, N.E. *et al.* The optimal measure of allelic association. *Proc. Natl Acad. Sci. USA* **98**, 5217–5221 (2001).
3. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
4. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
5. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232.
6. Rioux, J.D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn's disease. *Nature Genet.* **29**, 223–228.
7. Tishkoff, S.A. *et al.* Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. *Science* **271**, 1380–1387 (1996).
8. Kidd, K.K. *et al.* A global survey of haplotype frequencies and linkage disequilibrium at the *DRD2* locus. *Hum. Genet.* **103**, 211–227 (1998).
9. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
10. Bonnen, P.E. *et al.* Haplotypes at *ATM* identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 1437–1451 (2000).
11. Hollox, E.J. *et al.* Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* **68**, 160–172 (2001).
12. Rieder, M.J., Taylor, S.L., Clark, A.G. & Nickerson, D.A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22**, 59–62 (1999).
13. Harding, R.M. *et al.* Evidence for variable selective pressures at *MC1R*. *Am. J. Hum. Genet.* **66**, 1351–1361 (2000).
14. Fullerton, S.M. *et al.* Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).
15. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
16. Templeton, A.R., Weiss, K.M., Nickerson, D.A., Boerwinkle, E. & Sing, C.F. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* **156**, 1259–1275 (2000).
17. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222.
18. Martin, E.R. *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
19. Lucassen, A.M. *et al.* Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR. *Nature Genet.* **4**, 305–310 (1993).
20. Nickerson, D.A. *et al.* Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.* **10**, 1532–1545 (2000).
21. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
22. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
23. Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet.* **27**, 371–372 (2001).
24. Hugot, J.P. *et al.* Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
25. Longmate, J.A. Complexity and power in case-control association studies. *Am. J. Hum. Genet.* **68**, 1229–1237 (2001).
26. Drysdale, C.M. *et al.* Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
27. Mummidi, S. *et al.* Evolution of human and non-human primate CC chemokine receptor 5 gene and mRNA. Potential roles for haplotype and mRNA diversity, differential haplotype-specific transcriptional activity, and altered transcription factor binding to polymorphic nucleotides in the pathogenesis of HIV-1 and simian immunodeficiency virus. *J. Biol. Chem.* 275, 18946-18961 (2000).
28. Joosten, P.H., Toepoel, M., Mariman, E.C. & Van Zoelen, E.J. Promoter haplotype combinations of the platelet-derived growth factor α-receptor gene predispose to human neural tube defects. *Nature Genet.* **27**, 215–217 (2001).
29. Todd, J.A. *et al.* Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. *Nature* **338**, 587–589 (1989).
30. Degli-Esposti, M.A. *et al.* Ancestral haplotypes reveal the role of the central MHC in the immunogenetics of IDDM. *Immunogenetics* **36**, 345–356 (1992).
31. Farrall, M. *et al.* Fine-mapping of an ancestral recombination breakpoint in DCP1. *Nature Genet.* **23**, 270–271 (1999).
32. Mein, C.A. *et al.* Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res.* **10**, 330–343 (2000).