Am. J. Hum. Genet. 61:734–747, 1997

Association Mapping of Disease Loci, by Use of a Pooled DNA Genomic Screen

Lisa F. Barcellos,[1] William Klitz,[1] L. Leigh Field,[2] Rose Tobias,[2] Anne M. Bowcock,[3] Ross Wilson,[3] Mark P. Nelson,[1] Jane Nagatomi,[2] and Glenys Thomson[1]

[1]Department of Integrative Biology, University of California, Berkeley; [2]Department of Medical Genetics, University of Calgary, Calgary; and [3]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas

Summary

Genomic screening to map disease loci by association requires automation, pooling of DNA samples, and 3,000–6,000 highly polymorphic, evenly spaced microsatellite markers. Case-control samples can be used in an initial screen, followed by family-based data to confirm marker associations. Association mapping is relevant to genetic studies of complex diseases in which linkage analysis may be less effective and to cases in which multigenerational data are difficult to obtain, including rare or late-onset conditions and infectious diseases. The method can also be used effectively to follow up and confirm regions identified in linkage studies or to investigate candidate disease loci. Study designs can incorporate disease heterogeneity and interaction effects by appropriate subdivision of samples before screening. Here we report use of pooled DNA amplifications—the accurate determination of marker-disease associations for both case-control and nuclear family–based data—including application of correction methods for stutter artifact and preferential amplification. These issues, combined with a discussion of both statistical power and experimental design to define the necessary requirements for detecting of disease loci while virtually eliminating false positives, suggest the feasibility and efficiency of association mapping using pooled DNA screening.

Introduction

The rapid development and application of highly informative microsatellite markers (NIH/CEPH Collaborative Mapping Group 1992; Weissenbach et al. 1992; Gyapay et al. 1994; Murray et al. 1994; Dib et al. 1996) has greatly facilitated the localization of disease loci in >400 rare Mendelian disorders (Cooper and Schmidtke 1992; Gottesman and Collins 1994; McKusick 1994). Utilization of large multigenerational pedigrees with a number of affected individuals is a proved strategy for isolating chromosomal regions containing disease genes by use of microsatellite markers and LOD-score linkage analysis (Ott 1991; Lander and Schork 1994; Jorde 1995; Weeks and Lathrop 1995). Linkage studies are often followed by association studies to further localize the predisposing gene (Jorde et al. 1994; Lander and Schork 1994; Jorde 1995; Weeks and Lathrop 1995). Linkage disequilibrium, the basis of marker associations, is typically found within a distance of 500 kb, ~0.5 cM (Jorde et al. 1994). For the vast majority of diseases, disease genes mapped to date have shown linkage disequilibrium with markers sufficiently close to the disease gene—for example, cystic fibrosis, Huntington disease, Wilson disease, Batten disease, Friedreich ataxia, myotonic dystrophy, torsion dystonia, hemochromatosis, diastrophic dysplasia, adult-onset polycystic kidney disease, the familial breast cancer gene BRCA1 in the Ashkenazi Jewish population, and others (Weir 1989; Hanauer et al. 1990; Harley et al. 1991; Håstbacka et al. 1992; Ozelius et al. 1992; Pound et al. 1992; Skraastad et al. 1992; Jazwinska et al. 1993; Lerner et al. 1994; Bowcock et al. 1994; Freidman et al. 1995; Feder et al. 1996).

For many genetic diseases, including complex diseases that do not demonstrate simple Mendelian inheritance, large multigenerational pedigrees suitable for linkage analysis may not be available. Sampling and linkage analysis of nuclear families with pairs of affected siblings is another commonly used strategy. However, the detection of genetic factors in linkage studies for complex diseases such as bipolar disorder, schizophrenia, rheumatoid arthritis, insulin-dependent diabetes mellitus (IDDM), non–insulin-dependent (NIDDM) diabetes, and multiple sclerosis (MS), (Davies et al. 1994; Field et al. 1994, 1996; Hashimoto et al. 1994; Copeman et al. 1995; Owerbach and Gabbay 1995; Antonarakis et al. 1996; Cornélis et al. 1996; Ebers et al. 1996; Friemer et al. 1996; Hanis et al. 1996; Kuokkanen et al. 1996; Luo et al. 1996; Multiple Sclerosis Genetics Group 1996; Sawcer et al. 1996) has proved difficult. Few re-

Received July 24, 1996; accepted for publication June 10, 1997.

Address for correspondence and reprints: Dr. William Klitz, Department of Integrative Biology, 3060 Valley Life Sciences Building, University of California, Berkeley, CA 94720. E-mail: wklitz@allele5.biol.berkeley.edu

0002-9297/97/6103-0031$02.00

gions containing putative disease loci have been uncovered in these studies, and no genes have yet been identified. In addition, most of these regions await confirmation in independent studies, with the exception of IDDM, for which there is currently sufficient evidence, for 4 of the >10 identified disease regions, to justify fine-mapping efforts (Luo et al. 1996; Todd and Farrall 1996). Multiple loci, each contributing modestly to disease expression, are likely to be involved in these complex disorders and may be difficult to detect by use of standard linkage approaches.

Large-scale association studies of candidate genes by use of case-control or nuclear family–based sample collections may in fact be more effective than linkage analysis, as a strategy to identify disease loci of modest effect (Risch and Merikangas 1996). Recent theoretical analysis of nuclear family–based data demonstrates that, apart from type I error, an association uncovered by use of nuclear family–based data must implicate a disease-predisposing locus linked to the marker locus (Ott 1989; Knapp et al. 1993; Spielman et al. 1993; Thomson 1995*b;* Spielman and Ewens 1996). Although the careful matching of cases and controls probably eliminates most population-stratification problems that might result in the detection of false associations, family-based data remove these concerns.

When an appropriate isolated population of relatively recent origin can be employed for genetic analysis of disease, a few hundred markers would be sufficient to screen the genome, if it can be assumed that linkage disequilibrium extends to ⩾10 Mb (Håstbacka et al. 1992). This would be most effective for simple monogenic diseases, but such extensive disequilibrium would prove a hindrance to gene localization. More generally, for large, essentially panmictic groups, 3,000 (and preferably 6,000) highly polymorphic, evenly spaced markers, with a 1-cM (0.5 cM) average distance between markers, would be required for an initial disease-association genomic screen, compared with the typical 300 markers, spaced at 10 cM, used for linkage studies with family data. A screen of this resolution would ensure that any disease gene would be, on average, <0.5 cM (0.25 cM) from a marker. Currently, the number of polymorphic microsatellites totals >12,000; 9,000 of these have heterozygosities >.70 and are predominantly dinucleotide markers (Bowcock et al. 1996). Although this translates to an average map resolution of 273 kb (0.273 cM), there are still likely to be sizable gaps. However, as progress in physical mapping efforts continues, higher-density and more evenly spaced genetic maps are soon anticipated—for example, 100 kb (0.1 cM) by 1999 (Cox and Myers 1996).

Association genomic screening of this magnitude becomes feasible with the use of both pooled DNA samples and automation of microsatellite typing methods to reduce labor and reagent costs. DNA-pooling strategies have been successfully documented for homozygosity mapping of candidate regions in autosomal recessive disorders, using isolated populations (Carmi et al. 1995; Nystuen et al. 1996; Scott et al. 1996; Sheffield et al. 1996*a,* 1996*b*). Pooling individual DNA samples renders the determination of allele frequencies much more efficient, since the same information obtained from the analysis of a large number of individuals can be extracted from pooled data by use of just a few coamplifications. The recent development of a mathematical method for correction of stutter artifact (or shadow bands) that occur primarily in the amplification of dinucleotide microsatellites (Perlin et al. 1995) has also improved the accuracy of microsatellite allele-frequency estimates using pooled data. In addition, large numbers of polymorphic tri- and tetranucleotide markers, which are generally more robust and show less stutter, are now available (Sheffield et al. 1995). Conventional autoradiographic methods for microsatellite typing have been replaced with much more efficient automated fluorescence-based systems, which allow simultaneous analysis of multiple markers with overlapping sizes within one gel lane (Reed et al. 1994; Kobayashi et al. 1995). DNA pooling and automated typing has recently been used to screen inbred mice for obesity genes (Taylor and Phillips 1996).

We propose the use of pooled DNA amplifications of microsatellite markers to facilitate efficient, cost-effective, high-resolution genome screening for detection of disease loci by association. Examples of marker-disease association are demonstrated by use of three microsatellite markers and data from two diseases, including both case-control and nuclear family–based samples. The problems of preferential amplification and stutter artifact that are present for many microsatellites are addressed. Successfully implementing a DNA-typing scheme using pooled samples produces allele-frequency data on thousands of genetic markers. The effective analysis and interpretation of these data require application of a clearly defined theoretical and statistical scheme. A detailed statistical power table is presented to guide study design for association mapping of disease genes, along with example experimental designs and laboratory-protocol considerations. Our results suggest the enormous potential of genomic screening using pooled DNA, as a valuable tool in the effort to identify both disease and trait loci generally.

Material and Methods

DNA Samples

Caucasian hemochromatosis patient samples ($n = 51$) were provided through a collaboration with Dr. George Sensabaugh, at University of California, Berkeley, and

Dr. David Baer, at Kaiser Hospital, Oakland. Unrelated Caucasian control individuals ($n = 75$) were collected in northern California. IDDM-affected sib-pair families ($n = 100$) were obtained from the Human Biological Data Interchange.

DNA Quantitation

Samples were diluted to approximate concentrations of 50 ng/μl (on the basis of spectrophotometric readings) and were quantitated (in duplicate) by use of the 96-well-format PicoGreen™ fluorescent assay (Molecular Probes), according to the manufacturer's protocol. In brief, each DNA sample was added to 100 μl of 10^{-4} M Tris-EDTA, and 1 μl of PicoGreen™ quantitation reagent, by use of 96-well Nunc microtiter plates. A standard curve was made from dilutions by use of a genomic DNA sample of known concentration. Excitation, using a spectofluorometer (Molecular Dynamics) was at 480 nm, and the emission wavelength was 520 nm. After quantitation, DNA pools were constructed for PCR amplification as appropriate for each disease—for example, a pool of 51 hemochromatosis patients and a pool of 75 control individuals. In addition, four pools of 100 individuals each (mothers, fathers, and first and second affected child) were constructed from the 100 IDDM families. Each pool contained 100 ng of DNA from each individual, at a final concentration of 10 ng/μl.

Microsatellite Typing

The following microsatellite markers were used in this study (Genome Database accession numbers for primer sequences are in parentheses): D6S105 (060097), F13A1 (156948), TH (225009), and MBP (063680). Forward primers for each primer pair—D6S105, F13A1, and TH (Research Genetics), and MBP (Operon Technologies)—were labeled with 6-FAM or 6-HEX phosphoramidites. All primer pairs were first optimized by use of individual DNA samples, to determine the appropriate amount of template, Mg^{2+} concentration, and thermocycling parameters for amplification efficiency. Reaction volumes (15 μl) containing 5–10 ng of genomic DNA, 1.5 mM $MgCl_2$, 10 mM Tris, 50 mM KCl, 200 μM of each dNTP, 25 ng of each primer, and 0.5 U of *Taq* polymerase (Boehringer-Mannheim) were prepared on ice and were overlaid with mineral oil. Amplifications were performed by use of a Perkin-Elmer 9600 or MJR thermocyler, and parameters were as follows: 2 min at 94°C initial denaturation, then 28 cycles (1 min at 94°C, 1 min at 56°C, and 45 s at 72°C) followed by a final extension of 30 min at 60°C. After marker optimization, pooled DNA samples were amplified for each microsatellite by use of conditions described above for individual typing. Both individual and pooled samples were diluted (D6S105 1:5, MBP 1:10, and F13A1 and TH 1:15) (and were pooled to run simultaneously, when this was possible), and 2.0 μl was combined with 0.5 μl of internal lane size standard (TAMARA-350) and 3.5 μl of blue dextran loading buffer with formamide (Applied Biosystems). Electrophoresis was performed for individual or pooled alleles by use of 5%–6% denaturing (7M) acrylamide (National Diagnostics) on 36-cm (Well-to Read) plates at 3,600 V for 2.0–3.5 h, by use of a model 377A Sequencer (Applied Biosystems). After electrophoresis, analysis was performed by use of the GeneScan (version 2.0.1) software (Applied Biosystems) as described in the manufacturer's instruction manual. In brief, (*a*) gel lane tracking was checked manually, (*b*) The second-order least-squares sizing method was used to calculate peak sizes to .01 of a base, and (*c*) size-standard peaks were user defined.

Estimation of Allele Frequencies on the Basis of Pooled Amplifications

Peak heights derived from electropherograms of pooled (N individuals) DNA amplifications were converted to $2N$ allele-frequency counts for each pool size. The equivalent of the subtraction method for individual typing (Thomson 1995*b*) was applied to estimate the "affected family–based control" (AFBAC) population in the IDDM families (hereafter referred to as the "control population"), with family samples: for N nuclear simplex families the control population would be determined by subtracting the $2N$ allele-frequency counts in the pooled sample of one affected child from each family from the $4N$ parental allele counts of the pooled sample of both parents. N simplex families would yield $2N$ allele-frequency counts in patients and $2N$ allele-frequency counts in controls. In our N affected-sib-pair families, the two affected sibs in each family were analyzed separately, as if they came from simplex families; that is, there were two separate pools of patient DNA, and each family contributed one of their affected sibs to the first pool, the other to the second pool. The two patient pools were treated as independent samples (under the null hypothesis of no marker association), and the subtraction method was applied to both, yielding a total of $4N$ allele-frequency counts in patients and $4N$ allele-frequency counts in controls.

Stutter-Artifact Correction for Pooled Data

Stutter correction (Perlin et al. 1995) was applied to pooled allele frequencies for dinucleotide marker D6S105. An observed vector distribution $\boldsymbol{Y}$ of band peak heights is characterized by $\boldsymbol{Y} = \mathbf{A}X$, where the vector $\boldsymbol{X}$ denotes the true distribution of allele proportions and the rectangular matrix $\mathbf{A}$ consists of known stutter profiles for the alleles derived from individual typings at a locus. Individual typing results obtained from a random sample of 20 control individuals were used to construct

the allele-specific correction matrix for D6S105. Two rare alleles, 113 and 133, were not present in this sample, and matrix profiles for these two alleles were therefore determined by use of adjacent alleles. The pseudoinverse function in the program package "Mathematica" (Wolfram Research) was used to calculate the inverse of the rectangular matrix **A**; ***X*** was then determined from the product of $\mathbf{A}^{-1}$ and ***Y***.

Preferential-Amplification Correction for Pooled Data

The extent of preferential amplification was determined for each allele on the basis of individual typing results. A preferential-amplification factor can be defined as the ratio of peak heights, x/y, where x is the greater peak height (usually the smaller allele) and y is the lesser peak height. For example, if the smaller allele has a relative peak height of .62 and the larger allele has a height of .38, then the preferential-amplification factor for the smaller allele would be 1.63. In the pooled runs an observed peak height for such alleles was divided by the preferential-amplification factor before conversion of the peak heights to allele-frequency counts.

Statistics

Goodness-of-fit testing using the χ^2 statistic was used for all comparisons between pooled allele frequencies and those determined on the basis of individual typings. The P value from goodness-of-fit testing was used as a measure of the degree of closeness between the pooled and individually typed allele-frequency distributions; for example, a P value of .90 indicates very high concordance between the two distributions. The χ^2 test for heterogeneity was used to determine significance in contingency-table testing of all case-control and family data.

Power Calculations

Simulations of 100,000 or 1,000,000 runs each of case-control and simplex family–based data were performed by use of a range of association strengths, modes of inheritance, and disease-allele frequencies. The recombination fraction between the marker and disease loci was sufficiently small that it could be taken as zero. The general population was assumed to be random mating, and equations for family-based data were from Thomson (1995*b*). Parental marker alleles not transmitted to the affected child formed the control population. Power was determined by the proportion of times that the null hypothesis (H_o: no association of a marker locus with disease) is rejected for a given association strength. The α level, or probability that the null hypothesis will be rejected by chance alone, was investigated at five values: .0001, .001, .01, .05, and .10. The higher α levels, .05 and .10, become pertinent in multiphasic genome screens. Five levels of control allele frequency were studied: .1, .2, .3, .4, and .5.

Ideally, the goal is to identify all loci contributing genetic predisposition to a disease; however, it is necessary to restrict our attention to those loci whose influence is great enough for detection. The ability to detect actual disease-predisposing loci by association analysis depends on their genetic contribution to disease; association strength (absolute difference in frequency of an allele in patients and controls) can be used as a relative measure of genetic contribution in this context.

Results

DNA Pooling Using Microsatellites, with Corrections for Stutter and Preferential Amplification

Disease associations can be revealed by PCR amplification of microsatellite markers and comparison of pools of patient and ethnically matched control DNA samples. The locus responsible for abnormal iron loading in hemochromatosis, for example, shows an association with allele 121 of the dinucleotide marker D6S105, located within a region 3.3 cM telomeric to HLA-A on chromosome 6p21 (Raha-Chowdhury et al. 1995; Feder et al. 1996). Significant stutter-artifact peaks were observed for D6S105, and an allele-specific matrix was designed by use of the individual D6S105 genotyping profiles for correction of pooled data (see Material and Methods). Table 1 contains the stutter-correction matrix **A** for D6S105. The columns of matrix **A** represent the stutter profiles for each of the observed alleles derived from individual genotyping results. The distribution of fragments for allele 133, for example, is .68, .26 and .06, for fragment sizes 133, 131, and 129, respectively. The pseudoinverse of this rectangular matrix is multiplied by a vector ***Y***, the observed distribution of fragment sizes in the pooled samples, to give the allele-frequency estimates, vector ***X***. We give the example of the observed pooled DNA-control frequency profile ($2N = 150$) and the resulting allele-frequency estimates (table 1).

The allele distributions for D6S105, in both patient and control samples, were obtained from individual typing and were compared with estimated frequencies from pooled data. Results determined by use of a pool of 75 control individuals are shown in figure 1, where the P value for this comparison, .91 (vs. $P < 10^{-5}$, uncorrected for stutter), demonstrates the closeness of corrected pooled allele counts to individual typing results. This experiment also exemplifies the limits of the technique: The mean divergence between individual and pooled allele frequencies is 0.7%/allele and ranges from 0.1%/allele, for allele 121, to 3.1%/allele, for allele 123.

Comparison of a pool of 51 hemochromatosis patients with a pool of 75 control individuals typed for D6S105 (shown in fig. 2) revealed a striking positive association of allele 121 in the patients, accompanied

Table 1

D6S105 Stutter-Correction Matrix A, Observed Pooled Fragment Distribution of Controls, *Y*, and Resulting Estimated Allele Frequencies *X*

Peak Height from Automatic DNA Sequencer, for allele													
133	131	129	127	125	123	121	119	117	115	113	*Y*	*X*	Allele
.68	0	0	0	0	0	0	0	0	0	0	2.19	3.22	133
.26	.68	0	0	0	0	0	0	0	0	0	4.22	4.97	131
.06	.26	.55	0	0	0	0	0	0	0	0	9.33	14.26	129
0	.06	.27	.73	0	0	0	0	0	0	0	16.89	17.45	127
0	0	.18	.27	.68	0	0	0	0	0	0	47.56	59.24	125
0	0	0	0	.26	.73	0	0	0	0	0	26.13	13.75	123
0	0	0	0	.06	.27	.73	0	0	0	0	28.03	29.38	121
0	0	0	0	0	0	.23	.73	0	0	0	8.73	2.70	119
0	0	0	0	0	0	.04	.27	.73	0	0	4.98	3.94	117
0	0	0	0	0	0	0	0	.23	.80	0	1.94	1.33	115
0	0	0	0	0	0	0	0	.04	.20	.80	.00	.00	113
0	0	0	0	0	0	0	0	0	0	.20			

by deficits among most other alleles ($\chi^2 = 47.61$, df = 7, $P < 10^{-5}$, corrected for stutter), with values virtually indistinguishable from individual typing results ($\chi^2 = 47.95$, df = 7, $P < 10^{-5}$). Significant differences were also detected between pooled patient and control data when there was no mathematical correction for stutter artifact ($\chi^2 = 36.97$, df = 7, $P < 10^{-5}$), even though the determination of allele frequencies within each pooled sample was not as accurate. Multiple comparisons were made for D6S105 by use of independent amplifications of both patient and control pools, to demonstrate reproducibility of these results. This example shows that a dinucleotide microsatellite marker typed by use of pooled DNA samples is able to clearly signal the presence of a disease locus by use of association analysis in a case-control study design.

Preferential amplification, which may occur when the range in microsatellite allele sizes is sufficiently great, was explored with the tetranucleotide marker MBP. The smallest of the eight MBP alleles, 212 bp, consistently amplified 30% more than other alleles, in both individual genotyping profiles and pooled data, resulting in a preferential-amplification factor of 1.3. For this comparison of individually typed and pooled samples of 75 individuals, $P > .99$, compared with $P = .83$ for uncorrected data (not shown). No significant stutter artifacts or preferential amplification of smaller alleles were observed for the other two loci examined, F13A1 and TH. The methods of adjustment presented here for both dinucleotide stutter and preferential amplification provide promising solutions to these problems, and they need to be investigated further, for accuracy and reproducibility, by use of a large panel of di-, tri-, and tetranucleotide microsatellite markers.

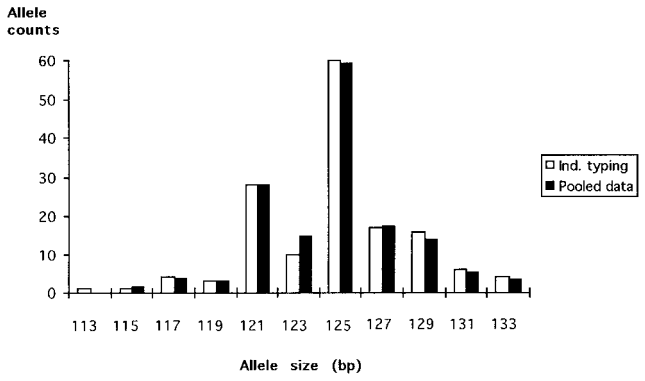

Figure 1 Pooled and individual typing results for D6S105 microsatellite and pool size 75. Pooled data were adjusted for stutter artifact by use of the matrix shown in figure 1 ($P = .91$, vs. 10^{-5} for unadjusted data).

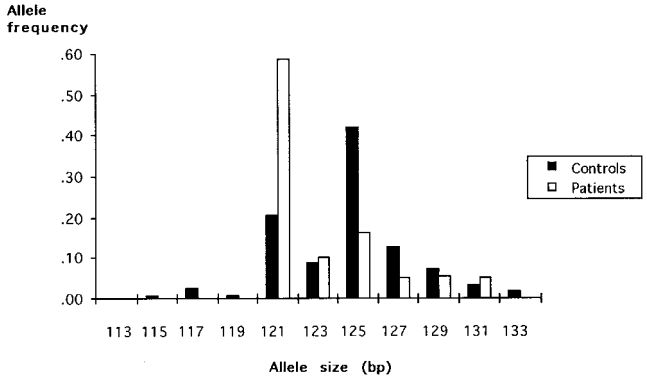

Figure 2 Hemochromatosis patients ($n = 51$) and control individuals ($n = 75$), typed by use of pooled samples, for D6S105 microsatellite. Data were adjusted for stutter artifact by use of the matrix shown in figure 1 ($\chi^2 = 47.61$, 7 df, $P < 10^{-5}$).

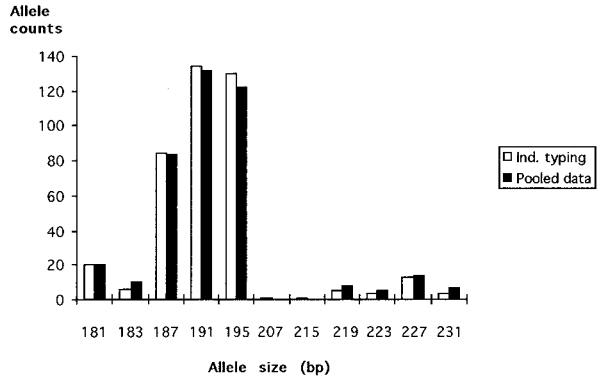

Figure 3 Pooled and individual typing results for F13A1 microsatellite AFBAC control population, determined by use of the subtraction method described by Thomson (1995*b*). A collection of 100 families each with two affected sibs were used ($P > .95$).

DNA Pooling and Marker-Disease Associations by Use of Nuclear Family–Based Data

A DNA-pooling strategy can also be used with nuclear family–based sample collections, to identify marker-disease associations. DNA samples obtained from mothers, fathers, and affected children are pooled separately for PCR amplification of a particular marker. The subtraction method (Thomson 1995*b*) can then be applied to pooled data to estimate a parental nontransmitted (i.e., AFBAC) control population for comparison with patient allele frequencies (see Material and Methods). A collection of 100 nuclear IDDM families with affected sib pairs was utilized to demonstrate the application of this method.

IDDM is strongly associated with HLA class II DR3 and DR4 haplotypes (IDDM1) on chromosome 6p21.3 (Karvonen et al. 1993), and a weaker genetic effect, due to IDDM2 which is located on chromosome 11p15, is well documented (Thomson et al. 1989; Bennett et al. 1995). Two tetranucleotide microsatellite markers—TH, located 8.4 kb upstream from IDDM2 on 11p15.5, and F13A1, located outside the HLA region on chromosome 6p24.2—were chosen for investigation, to serve as examples of loci having, respectively, a known weak disease association and no association.

Four pools of 100 individuals each (mothers, fathers, and first and second affected child) were constructed and amplified for both markers. The allele distribution of the pooled control population was calculated and compared with that obtained from individual typing results, by use of goodness-of-fit testing. The results for the F13A1 control alleles (fig. 3) demonstrate that the two populations are extremely similar ($P > .95$). Similar results were obtained for TH ($P > .90$). Multiple comparisons were made for both markers, by use of independent pooled amplifications and pool sizes ≤ 200 (data not shown), to demonstrate reproducibility of these results. These investigations show that control allele frequencies can be derived accurately from family data by use of pooled samples.

The allele distributions of the pooled control populations were compared with pooled patient distributions, for both markers, to test for disease association. As observed with individual typing, F13A1 showed no association with IDDM ($P > .10$). TH, however, was weakly associated with IDDM, in both individually typed ($\chi^2 = 9.61$, df = 5, $P < .05$) and pooled data sets ($\chi^2 = 10.63$, df = 4, $P < .05$); the largest effects were due to decreases in the patient alleles 191 and 198 (see fig. 4). This is due to linkage disequilibrium between certain alleles (particularly 191 and 198) of the TH locus and IDDM2 (VNTR) class III alleles (Bennett et al. 1995). IDDM2 class I alleles are associated with disease resulting both in significantly lower patient class III allele frequencies and in correspondingly lower frequencies of TH alleles 191 and 198. In this case, the pooling strategy is demonstrated to be effective in signaling the existence of a disease locus, through disequilibrium with a linked marker locus, by use of family data. This is in spite of a number of demanding circumstances, including the inability of conventional linkage analysis to detect the disease locus (Cox et al. 1988; Davies et al. 1994; Field and Nagatomi 1994), the presence of protective alleles, the association of the disease allele with two microsatellite allele markers, and the relatively weak overall influence of the disease locus, IDDM2 being secondary to HLA (IDDM1).

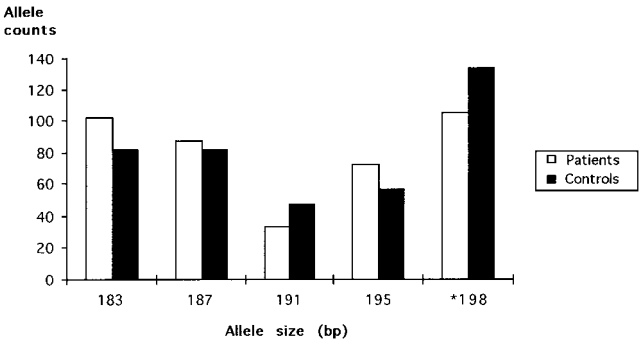

Figure 4 IDDM patients versus AFBAC controls (pooled data), with use of TH microsatellite marker. Significant heterogeneity is observed between the two allele distributions (χ^2 10.63, 4 df, $P < .05$). Of the seven known alleles at the TH locus, two (198 and 199) differ by only one base. Both alleles are easily distinguished in individual genotyping analyses but are detected only as a single peak (*) in a pooled DNA amplification; this did not, however, pose a problem for pooled-data comparisons involving TH in this study, since the sum of the allele frequencies for 198 and 199, determined by use of individual typing, was statistically indistinguishable from the combined allele frequency calculated by use of pooled data. In Caucasian populations, the 198 allele is much more common (frequency ~35%) than the 199 allele (frequency ~1%) (Puers et al. 1993).

Table 2

Statistical Power to Detect True Association

Association Strength and N[a]	Statistical Power to Detect True Association at α =[b]				
	.01%	.1%	1%	5%	10%
.20:					
1,000	100	100	100	100	100
500	100	100	100	100	100
400	100	100	100	100	100
300	100	100	100	100	100
200	97–100	99–100	100	100	100
100	58–89	79–97	93–99	98–100	100
50	15–36	34–62	58–85	83–96	90–98
.10:					
1,000	99–100	100	100	100	100
500	73–99	89–100	97–100	99–100	100
400	55–96	77–99	93–100	98–100	99–100
300	34–84	58–95	82–99	94–100	97–100
200	15–53	32–76	60–92	81–98	88–99
100	3–13	10–31	29–59	52–81	64–88
50	1–2	3–9	12–27	31–52	42–64
.05:					
1,000	24–82	45–93	72–99	89–100	94–100
500	5–30	15–54	37–79	62–92	73–96
400	3–19	10–40	28–67	51–86	65–92
300	1–10	6–25	20–52	42–75	54–84
200	1–4	3–12	12–33	30–57	40–69
100	. . .[c]	1–3	6–14	17–32	26–45
50	. . .[c]	. . .[c]	3–6	12–19	19–29

[a] "Association strength" refers to the true difference in marker allele frequencies in patients and controls—that is, at the whole-population level rather than at the sample level; for example, an association strength of .20 observed for a marker allele with frequencies of .7 and .5 in patients and controls, respectively, has an odds ratio of 2.3, whereas an association strength of .20 for marker allele frequencies of .3 and .1 has an odds ratio of 5.1. N = number of patients in a case-control study (N controls are assumed) or number of affected children in nuclear families ascertained for the presence of at least one affected child.

[b] α = Type I probability level used in detecting associations.

[c] Values too close to type I error to be considered.

Statistical-Power Considerations for Association Mapping

An appropriate experimental design for association disease mapping must address two issues: (1) power, the probability of detecting a true disease association, and (2) elimination of false associations, which represent type I error. These form the basis of several important considerations in the determination of the best experimental design for association mapping, given intrinsic properties of a disease, such as population prevalence, age at onset, and resources available to perform a genome screen. We utilized simulation to determine power for nuclear family–based and case-control data for a range of sample sizes and association strengths pertinent to a genome screen (table 2), where association strength refers to the true difference (at the population level—not at the sample level), in marker-allele frequencies, between patients and controls (population control and parental nontransmitted alleles, respectively, for case-control and nuclear family–based data). A biallelic marker system was studied, and independence of results for adjacent markers was assumed.

Power for association mapping is dependent on marker-allele frequencies and a number of unknown parameters: mode of inheritance of the disease, disease-allele frequencies, linkage disequilibrium between the marker and disease loci, and the recombination fraction between the marker and disease loci. However, with both case-control and family-based data and with random-mating assumptions, our simulations show that power is determined only by the patient and control marker-allele frequencies, in which the unknown parameters are subsumed. Power is lowest when the frequency of a disease-associated allele in the control population (f_c) is .50 and is highest when the control allele frequency is low. When in table 2 a range of power values is given, the lowest value corresponds to f_c = .5, and the highest value corresponds to f_c = .1; for example, in the case of N = 200, association strength = .1, and α = 1%, when f_c = .1, power is .92, and, when f_c = .5, power is .60.

Power for high (.20), moderate (.10), and low (.05) association strengths are given table 2. Low association-strength effects are detectable only in larger sample sizes. At an association strength of .05, the odds ratio is 1.7 and 1.2, with control frequencies of .1 and .5, respectively.

These power calculations serve only as a guide in the design of an experimental protocol. The actual power in any situation will depend on a large number of factors, including the number of marker alleles, the method of analysis (Terwilliger 1995; Schaid 1996), and the actual form of the association—that is, positive or negative, one marker associated versus multiple markers, etc. These biallelic calculations overestimate power for highly polymorphic microsatellites, but, with appropriate modification of methods of association analysis—for example, use of a likelihood approach (Terwilliger 1995)—the power may not be greatly overestimated. Even for moderate sample sizes—for example, 200 patients and 200 controls—power is very high if the association strength is $\geq$.10 and the type I α level is $\geq$1%. For larger sample sizes, which are feasible with a pooling strategy, power remains high at nearly all α values and association strengths. Power at α = 10% is high even with smaller sample sizes and association strengths; however, higher α values mean more type I errors, and an appropriate study design must balance these two factors.

Experimental Design for a High-Resolution Association Genome Screen

A study can be individually designed, taking account of the power calculations in table 2, the availability of

Table 3

Illustrative Experimental Designs for Pooled Association Genome Screen

	PHASE I (INDIVIDUAL)	PHASE II (INDIVIDUAL)	PHASE III (FAMILY)
	Change in Type I Errors		
Experiment A:			
Sample size[a]	200/200	200/200	100
Markers typed[b]	3,000	150 + *D*	8 + *D*
α Level	.05	.05	.05
Type I errors	150	8	.4
	Type II Errors (% of True Errors Missed)		
Association strength:			
.20	<1	<1	1
.10	10	19	54
.05	56	81	95
	PHASE I (INDIVIDUAL)		PHASE II (FAMILY)
	Change in Type I Errors		
Experiment B:			
Sample size[a]	1,000/1,000		500
Markers typed[b]	6,000		6 + *D*
α Level	.001		.05
Type I errors	6		.3
	Type II Errors (% of True Errors Missed)		
Association strength:			
.20	0		0
.10	0		.5
.05	31		47

[a] Number of affected and control individuals. A family with one affected child gives the equivalent of one affected and one control individual, whereas a family with an affected sib pair gives two affected and two controls in a double-simplex analysis.

[b] *D* true allelic associations are present.

[c] When a range of power values are given in table 2, the midpoint (.3) of the control allele frequency values is used.

case-control and family samples, and the expected type I errors following the high-resolution genome screen. It is important that, if additional samples of patients and controls or families are available, they be utilized to maximize power at that stage. Family-based data, including simplex and affected-sib-pair families, can be used at any phase of the screen. We illustrate two multiphase experimental designs, with the aim of high power and elimination of type I errors (tables 3 and 4). Marker-disease associations remaining at the end of the studies are detecting genetic influences on disease, with most type I errors eliminated.

In experiment A (see table 3), three phases of screening are employed, each with moderate sample sizes—200 patients and 200 controls in each of phases I and II and 100 simplex families in phase III. In phase I, an overall genome screen is performed on case-control pooled samples by use of 3,000 markers. Results from adjacent markers are assumed to be independent, given the marker density, which means that experimental-design considerations for phase II should take account both of the actual number of marker associations detected and of their respective map distributions in phase I, if they are much higher than the number of type I errors expected. In phase II, associations detected in phase I, at $\alpha = .05$ (150 of which are expected to be type I errors), are tested in an independent set of pooled cases and controls collected while phase I screening is in progress. Phase II reduces the number of type I errors from 150 to 8. In phase III, 100 simplex nuclear-family samples are typed for those $(8 + D)$ associations replicated in phase II. Significant associations ($\alpha = .05$) that remain at the end of phase III are very unlikely to be type I errors (0.4 type I errors remain). Also, the nuclear family–based typing ensures that, apart from type I error, remaining associations are due to markers linked to disease loci. The number of markers to be typed decreases dramatically after each phase.

Associations confirmed in phase III would be individually typed in all samples (case-control and family samples). With the use of family-based data in which parental marker alleles are *never* transmitted to the affected sib or sib pairs in multiplex ascertained families, the AFBAC control population would also be calculated (Thomson 1995*b*), giving an unbiased estimate of population control allele frequencies when the recombination fraction between the marker and disease loci is sufficiently small that it can be taken as zero. The individually typed data could then be analyzed for relative predispositional effects, genotype effects, modes of inheritance, and parent-specific effects, including maternal/fetal interactions (Thomson 1995*a*). Efforts to finely map the genetic regions to localize the putative disease gene could be conducted simultaneously.

In experiment A, virtually all markers with association strengths >.20 have been revealed by the end of phase II (table 3). For markers with association strengths ⩾.10, <19% are missed after phase II, whereas, for those at an association strength of .05, 81% are missed. At the end of phase III, for markers with an association strength ⩾.20, detection is virtually complete (1% missed), whereas just over half (54%) of loci with an influence of .10 are missed, and those having a smaller influence (.05) are almost entirely missed.

Experiment B examines 6,000 markers, with sample sizes of 1,000 cases and 1,000 controls in phase 1 and with a sample size of 500 nuclear families in phase II.

Table 4

Total Genetic Contribution to a Disease Identified from Experiments A and B, When a Total of 15 Predisposing Loci Are Assumed

Experiment and Association Strength	Actual No. of Loci	Genetic Contribution	Phase I		Phase II		Phase III	
			No. of Loci Not Excluded	Genetic Contribution Detectable	No. of Loci Not Excluded	Genetic Contribution Detectable	No. of Loci Not Excluded	Genetic Contribution Detectable
A:								
.20	5	1.00 (57%)	5	57%	5	57%	5	56%
.10	5	.50 (29%)	4.5	26%	4	23%	2	15%
.05	5	.25 (14%)	2.5	6%	1.25	3%	.4	1%
Total	15	1.75 (100%)	12.0 (162)[a]	89%	10.25 (18.25)[a]	83%	7.4 (7.8)[a]	72%
B:								
.20	5	1.00 (57%)	5	57	5	57		
.10	5	.50 (26%)	5	26	4	23		
.05	5	.25 (17%)	3.5	12	2.34	10		
Total	15	1.75 (100%)	13.5 (19.5)[a]	95	11.3 (11.6)[a]	90		

[a] Total number of loci nominally significant.

The large sample size of phase 1 permits a stringent α level (.001), which reduces the number of false associations to six after the phase I screen and virtually eliminates type I errors after the phase II examination of $6 + D$ nominal associations. The large sample sizes also result in a very low rate of type II errors for association strengths >10% (>99% of loci identified).

The two experiments can be further used to characterize the success of a genome screen for disease loci, in terms of both the number of loci and the proportion of overall genetic influence identified (table 4). Simplistically, we assume microsatellites highly correlated with 15 actual disease-predisposing loci are present in a genome screen, 5 each at three levels of association strength (.2, .1, and .05), giving a total relative genetic contribution of 1.75. In experiment A, after phase I, on the basis of the type II errors (table 3), 80% (12/15) of the total number of loci involved in disease have not been excluded, whereas, at the end of the marker screen in phase III, approximately half (7.4) of the total susceptibility loci have been identified. In terms of genetic influence, however, largely because of the nearly complete identification (99%) of the high-association-strength loci, 72%, of the genetic contribution to the disease has been uncovered. The same analysis for experiment B reveals virtually complete identification of markers having an association strength $\geq$.10, whereas nearly half of the markers with association strengths of .05–.10 are identified, (table 4), making 75% of all disease loci identified. The large numbers of cases typed in phases I and II of experiment B allow >90% of the genetic contribution to be uncovered.

Discussion

Theoretical and empirical studies of linkage disequilibrium, as well as the availability of a large number of highly polymorphic microsatellite markers across the genome, indicate that high-resolution genome screens to detect marker associations can serve as an effective method to uncover disease loci. Our demonstrated reliability of pooling DNA samples to accurately estimate allele frequencies in case-control and nuclear family-based samples, as well as our development of appropriate experimental designs and protocols, suggest the feasibility of genome screening to map disease genes by association. This method can also be used either to efficiently follow up and confirm candidate regions identified in independent studies or to investigate functionally significant disease loci.

The results presented here are relevant to all disease studies but, in particular, to complex genetic diseases in which the mode of inheritance is unknown and in which multiple loci of modest effect, incomplete penetrance, heterogeneity, and interaction effects may be involved. Association mapping is also an appropriate strategy to study the genetic component of infectious diseases—for example, mapping the genes involved in progression to disease in tuberculosis or AIDS, for which linkage studies are not feasible. A recent report by Risch and Merikangas (1996) suggests that, in the identification of disease loci for complex disorders, large-scale testing by association analysis may be more effective than linkage analysis.

The central requirement for association mapping is that linkage disequilibrium be able to detect any locus within the map intervals afforded by microsatellite densities. The extent of disequilibrium can vary widely across the human genome. In the HLA region, which is probably subject to strong balancing selection, disequilibrium extends over 3 Mb (Klitz et al. 1992). In the region of the hemochromatosis gene, disequilibrium de-

clines monotonically over a region of 2–3 Mb, approximately localizing the site of the disease locus (Raha-Chowdhury et al. 1995; Feder et al. 1996). On the basis of current knowledge, the typical extent of linkage disequilibrium across the genome falls in the range of a few hundred to several hundred kilobases (Jorde et al. 1994). However, there are exceptions, with closely linked loci showing low levels of linkage disequilibrium—for example, the TAP genes within the HLA region (Klitz et al. 1995). Similarly, recombinational hotspots—as found, for example, in the β-hemoglobin region—can eliminate interlocus associations in the space of just a few kilobases (Chakravarti et al. 1984). Telomeric regions have shown lower levels of disequilibrium than are seen in regions more centromeric (Watkins et al. 1994). Although a 3,000-microsatellite (1-cM) genomic screen should detect many disease genes, a much finer—for example, a 6,000-microsatellite (0.5-cM) screen—will be the preferred goal. Even then, some disease loci will be missed, because of the lack of linkage disequilibrium, either in that particular region or with the disease locus—for example, BRCA1 in non-Ashkenazi Jewish populations, in which each family carries a unique mutation.

The ideal population for association mapping for monogenic diseases will be isolated, have a narrow population base, and be sampled not too many generations removed from the time during which a disease-causing mutation has occurred (Lander and Schork 1994). Genetically isolated populations, which should be considered when they are available, have been used successfully to identify a number of disease-gene regions (Carmi et al. 1995; Nystuen et al. 1996; Scott et al. 1996; Sheffield et al. 1996*a,* 1996*b*). In these situations, regions of disequilibrium may extend up to 10–20 cM, and a much lower marker density can be utilized. However, these are not essential requirements, as is illustrated by the observation that most genetic diseases show linkage disequilibrium with closely linked markers, even though they do not satisfy these conditions.

Because no complex trait or disease has been even partly characterized in exact genetic terms, the paradigm accounting for the success of disease-gene mapping to date—namely, that one or a few genetic mutations cause disease—deserves reexamination before being applied to complex diseases. Complex diseases may have their genetic underpinning determined, in part, by variation at loci segregating at polymorphic frequencies in a population. This is the case with many HLA-associated diseases—for example, IDDM, MS, and rheumatoid arthritis—in which common class II DRB1 and/or DQB1 alleles are directly involved in the disease process. The non-HLA IDDM genes have individual genetic effects that are much less than that of HLA and that may also reflect modifying effects of common variants. Furthermore, common allelic variants involved in disease may be the product of selective pressure. This is the explanation for the similar allele frequencies, at such loci, among long-isolated populations. The APOE locus (Weiss 1993) and the HLA complex (Hedrick 1991) are two examples. Polymorphism at each of these loci has been implicated in disease, through association mapping. The important feature for association mapping of loci that have been subject to selection is that disequilibrium can be anticipated to occur at longer than typical map intervals—and also that disequilibrium will survive over longer periods of time (Thomson 1977)—making these loci easier to detect and identify in population samples not taken from strictly defined islolates.

Several technical issues pertinent to an association genome screen using pooled DNA must be addressed. Perhaps the most challenging aspect of this approach is the accurate interpretation of microsatellite allele frequencies obtained from pooled amplifications. Preferential amplification of smaller alleles and stutter artifact (particularly with dinucleotide markers) will undoubtedly create problems. Some of these difficulties may be overcome by modification of the amplification cycle (Smith et al. 1995). Furthermore, the efficiency and reliability of pooled DNA amplifications may also be improved by integrating an automated hot-start method into the protocol, to increase yield and target specificity (Birch 1996). The successful application of a matrix-correction method, the adjustments for preferential amplification, and pool sample-size experiments reported here are very encouraging, but they need to be pursued in further studies.

Stutter artifacts and preferential amplification of alleles are marker dependent and will vary. Each marker in a high-resolution genomic screen should be independently optimized by use of individual samples before the initiation of pooled DNA amplifications. Optimization runs could include the same 20 randomly selected control individuals (40 chromosomes), which would serve to measure allelic stutter or preferential amplification, for mathematical adjustment of pooled data. A sample of this size will yield useful information on population frequencies and marker heterozygosity (if unknown) and could also signal strong disequilibrium with adjacent markers. In repeated amplifications of pools of 20 individuals by use of the D6S105 microsatellite marker, allele distributions compared between pooled and individually typed results were extremely close (all $P > .95$). More generally, binomial expectations show that an allele occurring in control samples at a frequency of .025 would be missed 36% of the time in the sample of 40 chromosomes but that an allele with a frequency of .10 would be missed only 1.5% of the time. Even if a particular allele were not sampled in the optimization runs, it is important to remember that artifacts produced by

stutter or preferential amplification act without bias on both patient and control samples. If a marker association is strong enough, significant differences will be observed between pooled patient and control allele frequencies, even without adjustment for stutter artifact.

Some microsatellite markers will not be suitable for DNA pooling, because of either poor primer design and resultant amplification inefficiency or unpredictable stutter-artifact patterns (particularly in dinucleotide repeats), which might not allow for reliable mathematical adjustment to accurately determine allele frequencies by use of pooled data. If either the disequilibrium of a marker or the association strength of a disease locus to a disease were moderate or low, then the signal from such a marker could be buried in the noise of stutter artifact and preferential amplification. With the increasing availability of larger numbers of microsatellites, including additional tri- and tetranucleotide markers (Sheffield et al. 1995), which generally are more robust and show less stutter, another marker within the same region could be utilized, original primers could be redesigned, or, alternatively, the marker could be used if all samples were individually typed.

The practical laboratory definition of several thousand microsatellite markers for use in pooled genomic screens is a significant task, but it must only be completed once, to make the method generally available for association mapping of human disease and trait loci. An automated sequencer or scanner for detection of fluorescently labeled PCR products is the ideal instrument for association screening with pooled DNA samples, because it allows a large amount of information to be obtained from each gel. To reduce labor and cost, the most efficient scheme for completing a pooled screen, including the preliminary marker-optimization phase, would be a large collaborative study of several diseases by independent laboratories typing all diseases, with a division of microsatellite markers among laboratories.

The advantage of using pooled DNA samples is especially clear in terms of cost and labor when compared with individual typing. In a screen of 200 patients and 200 controls (with pool sizes of 100 individuals each), the number of gels necessary to investigate 3,000 microsatellite markers would be drastically reduced, from ~6,000 for individual typing, to 125 for the pooled screen. (Calculations for typing of 3,000 markers are based on the use of 32 lanes [ABI 377 automatic sequencer with three overlapping primer dyes], on duplicate runs on each gel, for the pooled amplifications, and on a total of six markers typed per lane.) Although careful quantitation of DNA samples and careful construction of pools are both necessary when pooled amplifications are used, this is performed just once for an entire screen and constitutes a small fraction of the actual typing effort. Optimization runs for microsatellite markers must be completed before amplification of samples, for either a pooled or an individually typed screen; however, in pooling, construction of matrices for stutter correction will also be needed for some markers.

Large-scale association genome screening using pooled DNA samples also has several experimental advantages. In the case of complex diseases, both disease heterogeneity and interaction effects can be examined simultaneously by appropriate subdivision of patient groups prior to pooling. It can be hypothesized that different loci in different individuals may be contributing to a single disease phenotype. This possibility is particularly amenable to study in complex diseases where, for example, HLA associations have been determined and can be used to differentiate patient population pools (Tiwari and Terasaki 1985; Nepom and Erlich 1991; Klitz et al. 1994). In addition to predisposing genetic components within a subgroup of a particular disease, factors such as age at onset, sex, or other clinical descriptives can also be used for categorization while at the same time maintaining use of large-sample numbers for increased statistical power. Complex diseases for which linkage-based studies so far have been inconclusive, such as affective disorder and multiple sclerosis, are prime candidates for an association analysis using a pooled DNA genomic screen.

Acknowledgments

We wish to thank Emmanuelle Genin, Leslie Louie, Jill Hollenbach, Jeanette Wallin, Hugh Salamon, Shannon McWeeney, Mark Grote, Michael Esposito, George Sensabaugh, and Lisa Calandro for their significant contributions to this work. This research was supported by NIH grants GM35326 and GM56688 (both to L.F.B., W.K., and G.T.) and AR43177 (to A.M.B.) and by Medical Research Council of Canada grant MT-7910 (to L.L.F.). L.L.F. is an Alberta Heritage Medical Scientist.

References

Antonarakis SE, Blouin JL, Curran M, Luebbert H, Kazazian HH, Dombroski B, Housman D, et al (1996) Linkage and sib-pair analysis reveal a potential schizophrenia susceptibility gene on chromosome 13q32. Am J Hum Genet Suppl 59:A210

Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, et al (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284–292

Birch DE (1996) Simplified hot start PCR. Nature 381:445–446

Bowcock AM, Chipperfield MA, Ceverha P, Yetman E, Phung A (1996) Report of the DNA committee. In: Cuticchia AJ, Chipperfield MA, Foster PA (eds) Human Gene Mapping

1995: a compendium. Johns Hopkins University Press, Baltimore, pp 1454–1468

Bowcock AM, Tomfohrde J, Weissenbach J, Bonne-Tamir B, St George-Hyslop P, Giagheddu M, Cavalli-Sforza LL, et al (1994) Refining the position of Wilson Disease by linkage disequilibrium with polymorphic microsatellites. Am J Hum Genet 54:79–87

Carmi R, Rokhlina T, Kwitek-Black AE, Elbedour K, Nishimura D, Stone E, Sheffield V (1995) Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. Hum Mol Genet 4:9–13

Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazaian HH (1984) Nonuniform recombination within the human β-globin gene cluster. Am J Hum Genet 36:1239–1258

Cooper DN, Schmidtke J (1992) Molecular genetic approaches to the analysis and diagnosis of human inherited disease: an overview. Ann Med 24:29–41

Copeman JB, Cucca F, Hearne CM, Cornall RJ, Reed PW, Ronningen KS, Undlien DE, et al (1995) Linkage disequilibrium mapping of type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-33. Nat Genet 9:80–85

Cornélis F, Fauré S, Martinez M, Prud'homme JF, Fritz P, Dib C, Alves H, et al (1996) Genome search in rheumatoid arthritis affected sib-pair families suggests 3 susceptibility loci. Am J Hum Genet Suppl 59:A215

Cox DR, Myers RM (1996) A map to the future. Nat Genet 12:117–118

Cox NJ, Barker L, Spielman RS (1988) Insulin-gene sharing in sib pairs with insulin-dependent diabetes mellitus: no evidence for linkage. Am J Hum Genet 42:167–172

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130–136

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Ebers GC, Kukay K, Bulman DE, Sadovnick AD, Rice G, Anderson C, Armstrong H, et al (1996) A full genome search in multiple sclerosis. Nat Genet 13:472–476

Feder JN, Gnirke A, Thomas W, Tsuchiahashi Z, Ruddy DA, Basava A, Dormishian F, et al (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399–408

Field LL, Nagatomi J (1994) Linkage analysis: inadequate for detecting susceptibility loci in complex disorders? Am J Hum Genet Suppl 55:A347

Field LL, Tobias R, Magnus T (1994) A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin dependent diabetes mellitus. Nat Genet 8:189–194

Field LL, Tobias R, Thomson G, Plons S (1996) Susceptibility to insulin diabetes maps to a locus (IDDM11) on human chromosome14q24.3-q31. Genomics 33:1–8

Friedman LS, Szabo CI, Ostermeyer EA, Dowd P, Butler L, Park T, Lee MK, et al (1995) Novel inherited mutations and variable expressivity of BRCA1 alleles, including the founder mutation 185delAG in Ashkenazi Jewish families. Am J Hum Genet 57:1284–1297

Freimer NB, Reus VI, Escamilla M, Spesny M, Smith L, Service S, Gallegos A, et al (1996) An approach to investigating linkage for bipolar disorder using large Costa Rican pedigrees. Am J Med Genet 67:254–263

Gottesman M, Collins FS (1994) The role of the human genome project in disease prevention. Prev Med 23:591–594

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al (1994) The 1993–94 Généthon human genetic linkage map. Nat Genet 7:246–249

Hanauer A, Chery M, Fujita R, Driesel AJ, Gilgenkrantz S, Mandel JL (1990) The Friedreich ataxia gene is assigned to chromosome 9q13-q21 by mapping of tightly linked markers and shows linkage disequilibrium with D9515. Am J Hum Genet 46:133–137

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, et al (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13:161–166

Harley HG, Brook JD, Floyd J, Rundle SA, Crow S, Walsh KV, Thibault M-C, et al (1991) Detection of linkage disequilibrium between the myotonic dystrophy locus and a new polymorphic DNA marker. Am J Hum Genet 49:68–75

Hashimoto L, Habita C, Beressi JP, Delepine M, Besso C, Cambon-Thomsen A, Deschamp I, et al (1994) Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 371:161–164

Håstbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hedrick PW (1991) Evolutionary genetics of the major histocompatibility complex. Am Nat 143:945–964

Jazwinska EC, Lee SC, Webb SI, Halliday JW, Powell LW (1993) Localization of the hemochromatosis gene close to D6S105. Am J Hum Genet 53:347–352

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Karvonen M, Tuomilehto J, Libman I, LaPorte (1993) A review of the recent epidemiological data on the worldwide incidence of type 1 (insulin-dependent) diabetes mellitus. Diabetologica 36:883–892

Klitz W, Aldrich CL, Fildes N, Horning SJ, Begovich AB (1994) Localization of predisposition to Hodgkin disease in the HLA class II region. Am J Hum Genet 54:497–505

Klitz W, Stephens JC, Grote M, Carrington M (1995) Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the HLA class II region. Am J Hum Genet 57:1436–1444

Klitz W, Thomson G, Borot N, Cambon-Thomsen A (1992) Evolutionary and population perspectives of the human HLA complex. Evol Biol 26:35–72

Knapp M, Seuchter SA, Baur MP (1993) The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085–1093

Kobayashi H, Matise TC, Perlin MW, Marks HG, Hoffman

EP (1995) Towards fully automated genotyping: use of an X linked recessive spastic paraplegia family to test alternative analysis methods. Hum Genet 95:483–490

Kuokkanen S, Sundvall M, Terwilliger JD, Tienari PJ, Wikstrom J, Holmdahl R, Pettersson U, et al (1996) A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2. Nat Genet 13:477–480

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lerner TJ, Boustany R-MN, MacCormack K, Gleitsman J, Schlumpf K, Breakefield XO, Gusella JF, et al (1994) Linkage disequilibrium between the juvenile neuronal ceroid lipofuscinosis gene and marker loci on chromosome 16p12.1. Am J Hum Genet 54:88–94

Luo DF, Buzzetti R, Rotter J, Maclaren NK, Nistico L, Giovanni C, Pozzilli P, et al (1996) Confirmation of three susceptibility genes to insulin dependent diabetes mellitus: IDDM4, IDDM5 and IDDM8. Hum Mol Genet 5:693–698

McKusick VA (1994) Mendelian inheritance in man: a catalog of human genes and genetic disorders. Johns Hopkins University Press, Baltimore

Multiple Sclerosis Genetics Group (1996) A complete genomic screen for multiple sclerosis underscores a role for the major histocompatibility complex. Nat Genet 13:469–471

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillep J, et al (1994) A comprehensive human linkage map with centimorgan density. Science 265:2049–2054

Nepom GT, Erlich H (1991) MHC class-II molecules and autoimmunity. Annu Rev Immunol 9:493–525

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258:67–86

Nystuen A, Benke PJ, Merren J, Stone EM, Sheffield V (1996) A cerebellar ataxia locus identified by DNA pooling to search for linkage disequilibrium in an isolated population from the Cayman Islands. Hum Mol Genet 5:525–531

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

——— (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore

Owerbach D, Gabbay KH (1995) The HOXD8 locus (2q31) is linked to type I diabetes interaction with chromosome 6 and 11 disease susceptibility genes. Diabetes 44:132–136

Ozelius LJ, Kramer PL, de Leon D, Risch N, Bressman SB, Schuback DE, Brin MF, et al (1992) Strong allelic association between the torsion dystonia gene (DYT1) and loci on chromosome 9q34 in Ashkenazi Jews. Am J Hum Genet 50: 619–628

Perlin MW, Lancia G, Ng S-K (1995) Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am J Hum Genet 57:1199–1210

Pound SE, Carothers AD, Pignatelli PM, Macnicol AM, Watson ML, Wright AF (1992) Evidence for linkage disequilibrium between D16S94 and the adult onset polycystic kidney disease (PKD1) gene. J Med Genet 29:247–248

Puers C, Hammond HA, Jin L, Caskey CT, Schumm JW (1993) Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01 $(AATG)_n$ and reassignment of alleles in population analysis by using a locus-specific allelic ladder. Am J Hum Genet 53: 953–958

Raha-Chowdhury R, Bowen DJ, Stone C, Pointon JJ, Terwilliger JD, Shearman JD, Robson KJH, et al (1995) New polymorphic microsatellite markers place the haemochromatosis gene telomeric to D6S105. Hum Mol Genet 4:1869–1874

Reed PW, Davies JL, Copeman JB, Bennett ST, Palmer SM, Pritchard LE, Gough SCL, et al (1994) Chromosome-specific microsatellite sets for fluorescence based, semi-automated genome mapping. Nat Genet 7:390–395

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, Robertson N, et al (1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. Nat Genet 13:464–468

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Scott DA, Carmi R, Elbedour K, Yosefsberg S, Stone EM, Sheffield VC (1996) An autosomal recessive nonsyndromic-hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds. Am J Hum Genet 59:385–391

Sheffield VC, Kraiem Z, Beck JC, Nishimura D, Stone EM, Salameh M, Sadeh O, et al (1996*a*) Pendred syndrome maps to chromosome 7q21-34 and is caused by an intrinsic defect in thryroid iodine organification. Nat Genet 12: 424–426

Sheffield VC, Nishimura DY, Stone EM (1996*b*) Novel approaches to linkage mapping. Curr Opin Genet Dev 5:335–341

Sheffield VC, Weber JL, Buetow KH, Murray JC, Evan DA, Wiles K, Gastier JM, et al (1995) A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. Hum Mol Genet 4:1837–1844

Skraastad MI, Van de Vosse E, Belfroid R, Hold K, Vegter-van der Vlis M, Sandkuijl LA, Bakker E, et al (1992) Significant linkage disequilibrium between the Huntington disease gene and the loci D4S10 and D4S95 in the Dutch population. Am J Hum Genet 51:730–735

Smith JR, Carpten JD, Brownstein MJ, Ghosh S, Magnusson VL, Gilbert DA, Trent JM, et al (1995) Approach to genotyping errors caused by nontemplated nucleotide addition by Taq DNA polymerase. Genome Res 1:312–317

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Taylor BA, Phillips SJ (1996) Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling. Genomics 34:389–398

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

Thomson G (1977) The effect of a selected locus on linked neutral loci. Genet 85:753–788

——— (1995*a*) Analysis of complex genetic traits: an ordered-notation method and new tests for mode of inheritance. Am J Hum Genet 57:474–486

——— (1995*b*) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Thomson G, Robinson WP, Kuhner MK, Joe S, Klitz W (1989) HLA and insulin gene associations with IDDM. Genet Epidemiol 6:155–160

Tiwari JL, Terasaki P (1985) HLA and disease associations. Springer Verlag, New York

Todd JA, Farral M (1996) Panning for gold: genome-wide scanning for linkage in type I diabetes. Hum Mol Genet 5: 1443–1448

Watkins WS, Zenger R, O'Brien E, Nyman D, Eriksson AW, Renlund M, Jorde LB, et al (1994) Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. Am J Hum Genet 55: 348–355

Weeks DE, Lathrop GM (1995) Polygenic disease: methods for mapping complex disease traits. Trends Genet 11:513–519

Weir BS (1989) Locating the cystic fibrosis gene on the basis of linkage disequilibrium with markers. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (Eds) Multipoint mapping and linkage based upon affected pedigree members: Genetic Analysis Workshop 6. Alan R Liss, New York, pp 81–86

Weiss KM (1993) Genetic variation and human disease: principles and evolutionary approaches. Cambridge University Press, Cambridge, pp 180–204

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al (1992) A second-generation linkage map of the human genome. Nature 359:794–801