Am. J. Hum. Genet. 67:383–394, 2000

SNPing Away at Complex Diseases: Analysis of Single-Nucleotide Polymorphisms around *APOE* in Alzheimer Disease

Eden R. Martin,[1,2] Eric H. Lai,[3] John R. Gilbert,[1,2] Allison R. Rogala,[1,2] A. J. Afshari,[3] John Riley,[4] K. L. Finch,[5] J. F. Stevens[5] K. J. Livak,[5] Brandon D. Slotterbeck,[1,2] Susan H. Slifer,[1,2] Liling L. Warren,[1,2] P. Michael Conneally,[6] Donald E. Schmechel,[1,7] Ian Purvis,[4] Margaret A. Pericak-Vance,[1,2] Allen D. Roses,[1,3] and Jeffery M. Vance[1,2]

[1]Department of Medicine, [2]Center for Human Genetics, and [3]Bryan Alzheimer Disease Research Center, Duke University Medical Center, Durham, Duke University Medical Center, Durham, NC; [4]Glaxo Wellcome, Inc., Research Triangle Park, NC; [5]Glaxo Wellcome, Inc., London; [6]PE Biosystems, Foster City, CA; and [7]Department of Medicine and Molecular Genetics, Indiana University School of Medicine, Indianapolis

There has been great interest in the prospects of using single-nucleotide polymorphisms (SNPs) in the search for complex disease genes, and several initiatives devoted to the identification and mapping of SNPs throughout the human genome are currently underway. However, actual data investigating the use of SNPs for identification of complex disease genes are scarce. To begin to look at issues surrounding the use of SNPs in complex disease studies, we have initiated a collaborative SNP mapping study around *APOE*, the well-established susceptibility gene for late-onset Alzheimer disease (AD). Sixty SNPs in a 1.5-Mb region surrounding *APOE* were genotyped in samples of unrelated cases of AD, in controls, and in families with AD. Standard tests were conducted to look for association of SNP alleles with AD, in cases and controls. We also used family-based association analyses, including recently developed methods to look for haplotype association. Evidence of association ($P \leq .05$) was identified for 7 of 13 SNPs, including the APOE-4 polymorphism, spanning 40 kb on either side of *APOE*. As expected, very strong evidence for association with AD was seen for the APOE-4 polymorphism, as well as for two other SNPs that lie <16 kb from *APOE*. Haplotype analysis using family data increased significance over that seen in single-locus tests for some of the markers, and, for these data, improved localization of the gene. Our results demonstrate that associations can be detected at SNPs near a complex disease gene. We found that a high density of markers will be necessary in order to have a good chance of including SNPs with detectable levels of allelic association with the disease mutation, and statistical analysis based on haplotypes can provide additional information with respect to tests of significance and fine localization of complex disease genes.

Introduction

As human genetics moves toward identification of genes contributing to the susceptibility to common disorders and pharmacogenetic interactions, the mapping of traits has incorporated more techniques relying on genetic associations. Methods based on marker/disease associations can provide powerful tools for identification of disease loci. For localization of genes, association-based methods can be more powerful than linkage analysis, particularly when the contribution of these genes to the disease is small, as would be expected for complex diseases (Risch and Merikangas 1996). However, in order for association analysis to be useful, a dense map of markers will be required, since associations can generally only be found over small distances.

Recently, there has been great emphasis focused on identification of single-nucleotide polymorphisms (SNPs) in the human genome. It is anticipated that, during the next 3 years, ⩾100,000–200,000 SNPs will be identified by the Human Genome Project (Collins et al. 1998). Through a separate initiative, The SNP Consortium has been formed, with the goal of generating 300,000 SNPs during the next 2 years (Marshall 1999). SNPs have been postulated to be useful tools for identification of complex disease genes through association studies and, as such, are the next wave of markers for use in genetic analysis. It is estimated that SNPs occur, on average, every 1,000 bp and have a low mutation rate, both of which are characteristics that may have particular advantages for association analysis. However, as biallelic markers, SNPs are generally less informative than are microsatellite markers, which may be a disadvantage for both association and linkage analyses.

Although great progress already has been made to-

Received March 20, 2000; accepted for publication May 26, 2000; electronically published June 21, 2000.

Address for correspondence and reprints: Dr. Eden R. Martin, Duke University Medical Center, Box 3468, Durham, NC 27710. E-mail: emartin@chg.mc.duke.edu

0002-9297/2000/6702-0016$02.00

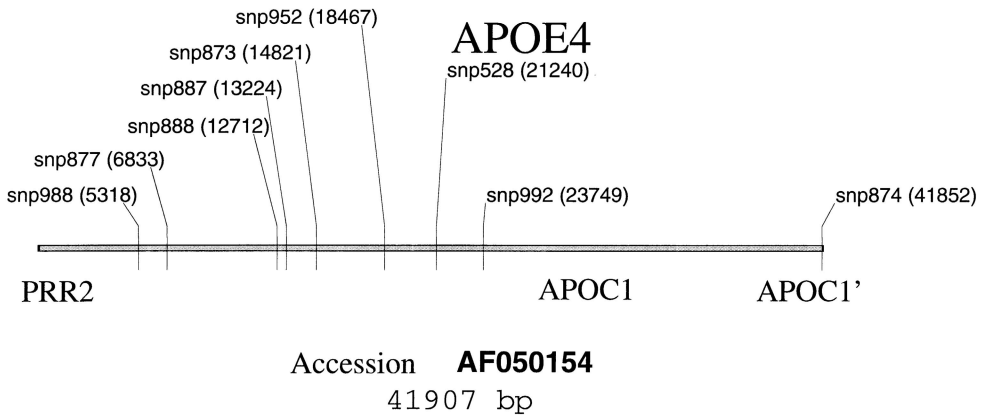

Figure 1 Physical map of SNPs on cosmid F19374 containing *APOE*

ward SNP development, several methodological questions remain to be addressed, in order to make optimal use of this growing resource of SNP markers. What density of SNPs should be used in analyses? Should SNPs be chosen specifically in genes or distributed randomly? Does knowledge of regional linkage disequilibrium help in the analysis? Is the frequency of the alleles of practical importance? What statistical methodology is most useful in the search for associations? What effect will the large number of multiple statistical effects have on the analysis? Can information from multiple SNPs be combined to improve the ability to detect and fine-map disease loci?

To begin to look at these questions, we have initiated a collaborative SNP-mapping study around *APOE* (MIM 107741), the well-established susceptibility gene for late-onset Alzheimer disease (AD) (MIM 104300). It is important to realize that this gene was identified using both traditional linkage analysis (Pericak-Vance et al. 1991) and association analysis (Corder et al. 1993; Saunders et al. 1993; Strittmatter et al. 1993*a*). Studies of *APOE* in primates and other mammals have suggested that APOE-4, the allele associated with AD, is the ancestral allele in humans (Hanlon and Rubinsztein 1995; Gerdes et al. 1996), and therefore it is expected that linkage disequilibrium will extend only over small distances. Thus, the APOE-4 allele should provide a rigorous challenge for gene identification through association analysis. Furthermore, there is a complex relationship between the APOE-4 genotype and disease status, with the APOE-4 allele being neither necessary nor sufficient for AD but, instead, modulating the risk for development of AD (Corder et al. 1993, 1994). This should make the results of this experiment especially useful in the definition of parameters and techniques for the utilization of SNPs in the mapping of common complex diseases. An initial analysis of 10 SNPs around *APOE* has suggested that a limited region of association surrounds the *APOE* gene (Vance et al. 1998; Martin et al. 2000). Here we describe additional analysis of 60 SNPs surrounding *APOE,* conducted to further investigate the model.

Subjects and Methods

Subjects

The case-control sample was composed of 220 cases with AD in which the age at onset was >59 years, mostly isolated (sporadic) cases collected at the Bryan Alzheimer's Disease Research Center (Bryan ADRC) at the Duke University Medical Center , and 220 controls in the same age group, most of whom were spouses of patients with AD. A small number of controls were spouses of known non-AD dementia patients collected at the Bryan ADRC. To improve homogeneity, all individuals included in this study were white. The family sample was composed of 184 families from the National Institute of Mental Health AD Genetics Initiative and the Indiana University AD Cell Repository (Blacker et al. 1997; Pericak-Vance et al. 1997). Of these 184 families, 92 contained phenotypically discordant sib pairs (one affected and one unaffected sibling), 60 contained at least two affected siblings and one unaffected sibling, and 32 had no unaffected siblings but contained at least one affected sib pair. This study focused on late-onset AD; thus, we required that the age at onset in each affected individual be >59 years. For the family data, we required that the age at onset be >59 years for all sampled affected family members, even if they were not selected from our database for this study.

SNP Generation and Mapping

We have previously described the generation of these SNPs (Lai et al. 1998) by using two different methods: YAC truncation and random sequencing. Mapping of

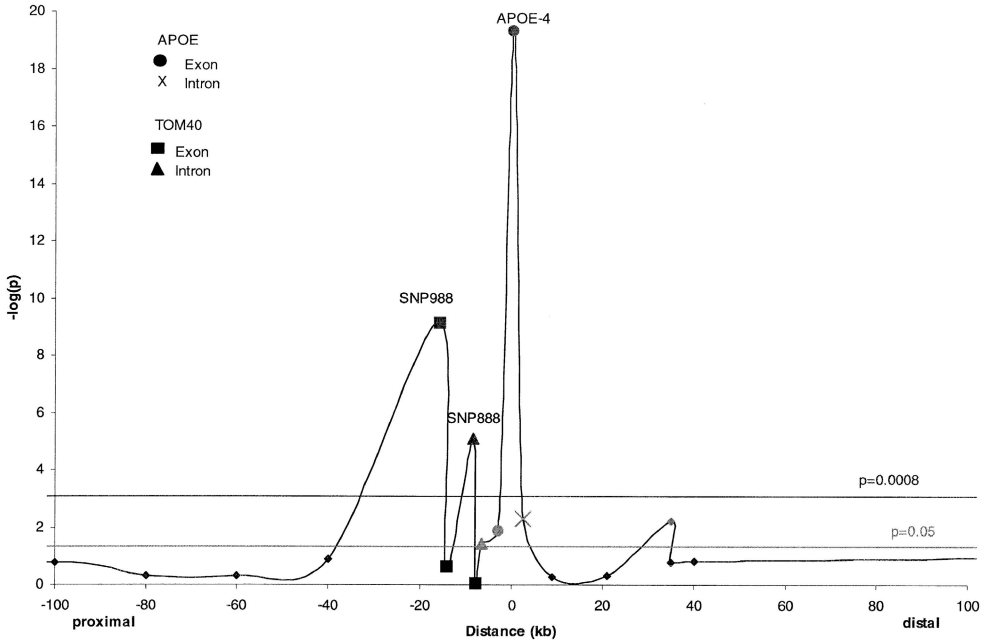

Figure 2 Plot of minus log of *P* value for case-control test for allelic association with AD, for SNPs immediately surrounding *APOE* (<100 kb).

the majority of the markers used the TNG4 radiation-hybrid panel. Additional mapping information came from restriction-enzyme digestion and subsequent hybridization of selected probes on cosmids containing *APOE,* as well as from published maps (Freitas et al. 1998) and available cosmid sequences from GenBank.

Genotyping

Detection was performed by TaqMan™ assay in 96-well plates after optimization for each primer set (Heid et al. 1996). This technique uses two probes, each complementary to one of the SNP alleles. Each probe contains a different-colored tag, whose fluorescent activity is prevented by the close presence of a "quencher" molecule, located on the other end of the probe. During PCR, the probe hybridizes downstream from the PCR primer, at the SNP site, creating a small double-strand segment. The 5′-3′ exonucleolytic activity of *Taq* DNA polymerase cleaves the probe as it replicates the template, releasing the fragmented probe into the solution, thus separating the quencher from the fluorescent tag. This allows the tag to excite in the presence of a laser, releasing the genotype-identifying color. DNA samples were quantitated twice, in duplicate, by PicoGreen fluorescent dye (Molecular Probes), and volumes were adjusted to ensure uniform concentration. A Hydra 96-well multipipetter (Robbins) was used to distribute 20 ng of DNA to each well. Samples of DNA from the Centre d'Étude du Polymorphisme Humain, which were sequenced for each SNP, were used as controls for each genotype. The DNA was allowed to dry, and 25 μl of master mix (8% glycerol, 1 × TaqMan™ buffer A, 5 mM $MgCl_2$, 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 400 μM dUTP, 0.05 U of Ampli*Taq* Gold DNA polymerase/μl, and 0.01 U of primer and probe/μl) was dispensed to each well, by a MultiProbe 204DT (Packard Instruments). The reaction was run on a GeneAmp PCR System 9700 (50°C for 2 min, 95°C for 10 min, and 95°C for 15 s; and 62°C for 1 min, for 35 cycles) and was read on an ABI Prism 7200 sequence detector.

Since the subjects in this study had minimal or no family structure, samples for six individuals were duplicated for each 96-well plate, to help detect potential loading and reading errors. This is a slight modification of the normal quality-control measures used by the Duke Center for Human Genetics in the genotyping of complex disorders (Rimmler et al. 1998). Technicians performing and reading the polymorphisms were blinded to the location of the duplicated samples. quality-control samples were compared in the Duke Center for Human Genetics Data Coordinating Center. Tests for deviation from Hardy-Weinberg equilibrium (HWE) also were conducted for all loci.

Statistical Analyses

Association analyses were conducted on the samples of cases and controls and in the family sample. In the case-control sample, allele and genotype frequencies were compared between case and control groups, by standard χ^2 tests for equality of proportions. To test for

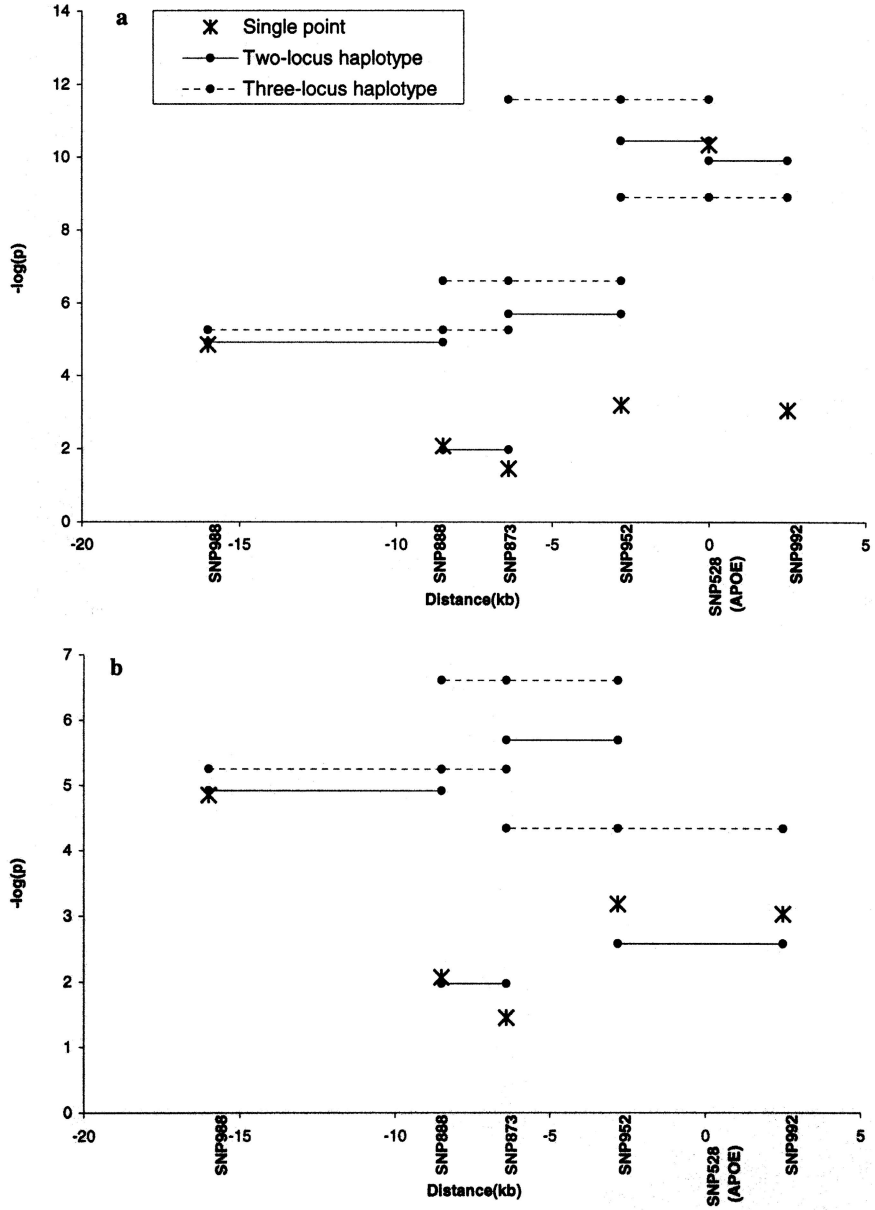

Figure 3 Plots of minus log of *P* value for TRANSMIT single-point and haplotype analysis of association at six SNPs in *APOE* region, including (*a*) and excluding (*b*) APOE-4 (SNP528).

association in the family sample, we used the weighted sibship disequilibrium test (WSDT) (Martin et al. 1999), which provides a valid test for association in sibships of arbitrary size. The WSDT is based on the difference between the number of times that a particular marker allele occurs in affected individuals and the number of times that the allele occurs in their unaffected siblings. χ^2 Approximations were used to calculate *P* values. Since 32 families had only affected siblings, thereby lacking the unaffected sibling required for the sibship tests described above, we also used a likelihood-based method from the program TRANSMIT (Clayton 1999), which does not require unaffected siblings. This method also was used to look for excess transmission of SNP haplotypes to affected individuals, providing an analysis of haplotype association, in contrast to the single-locus tests. There were nine extended pedigrees in our sample. For each of these, a single nuclear family was sampled, without respect to genotypes, for analysis in TRANSMIT. The WSDT remains a valid test of linkage disequilibrium, even in extended pedigrees.

Tests for allelic associations between pairs of markers

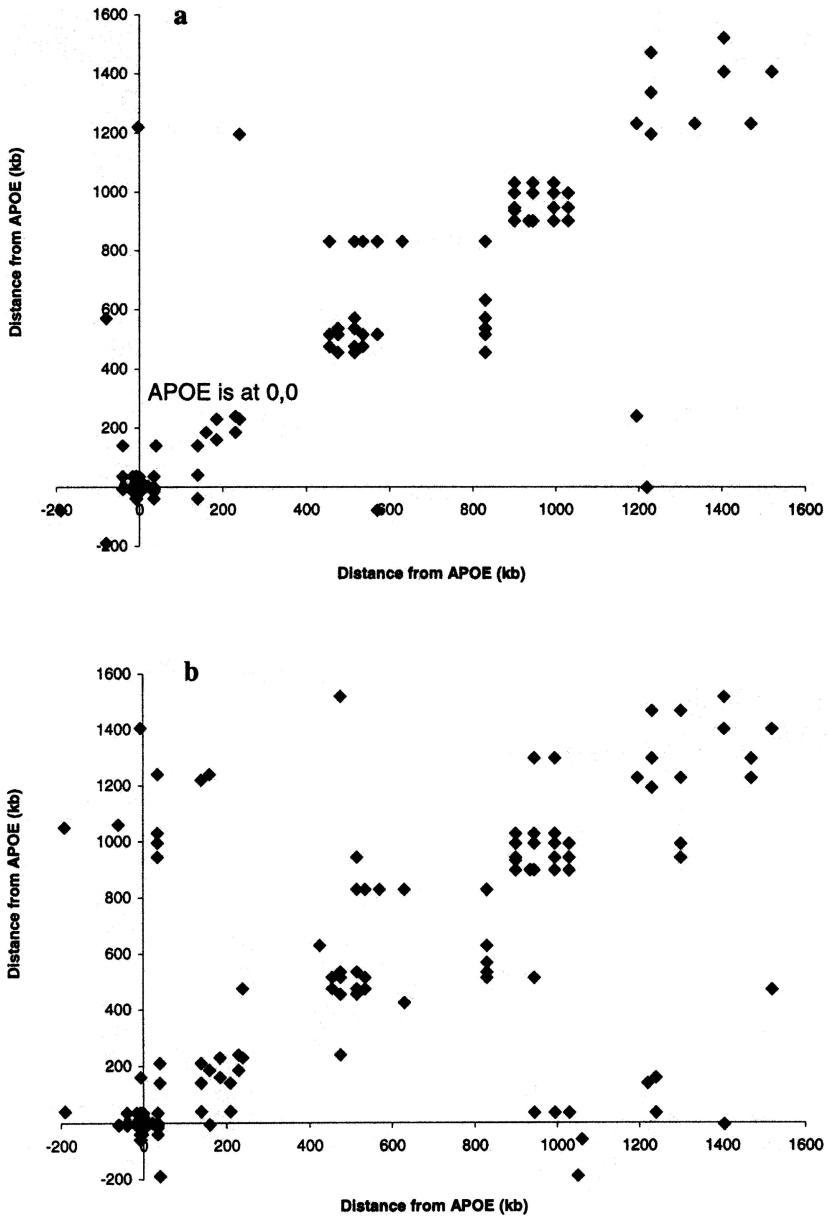

Figure 4 Plots of location of marker pairs with significant ($P < .01$) allelic association in controls (*a*) and cases (*b*)

were conducted in the case and control samples separately, by the GDA (genetic data analysis) program (GDA: Software for the Analysis of Discrete Genetic Data). We used an exact test based on the multinomial probability of the multilocus genotype, conditional on the single-locus genotypes (Zaykin et al. 1995). Significance was assessed by Monte Carlo simulation, by permuting the single-locus genotypes among individuals in the sample to simulate the null distribution. For each pair of SNPs, 3,200 replicate samples were simulated, to estimate the empirical *P* value. Since the single-locus genotypes are preserved in the Monte Carlo method, the test is not sensitive to deviations due to Hardy-Weinberg disequilibrium.

Two-point and multipoint linkage analyses were conducted by use of the package SIBLINK (Hauser and Boehnke 1998). Allele frequencies for input were estimated from the family data. Distances between markers for multipoint analyses were approximated on the basis of physical distance, with 1 Mb = 1 cM being assumed, with a minimum map distance of 0.001 cM between adjacent markers; although distance is known to be variable across the genome, comparison of the chromosome 19 physical map from Lawrence Livermore National

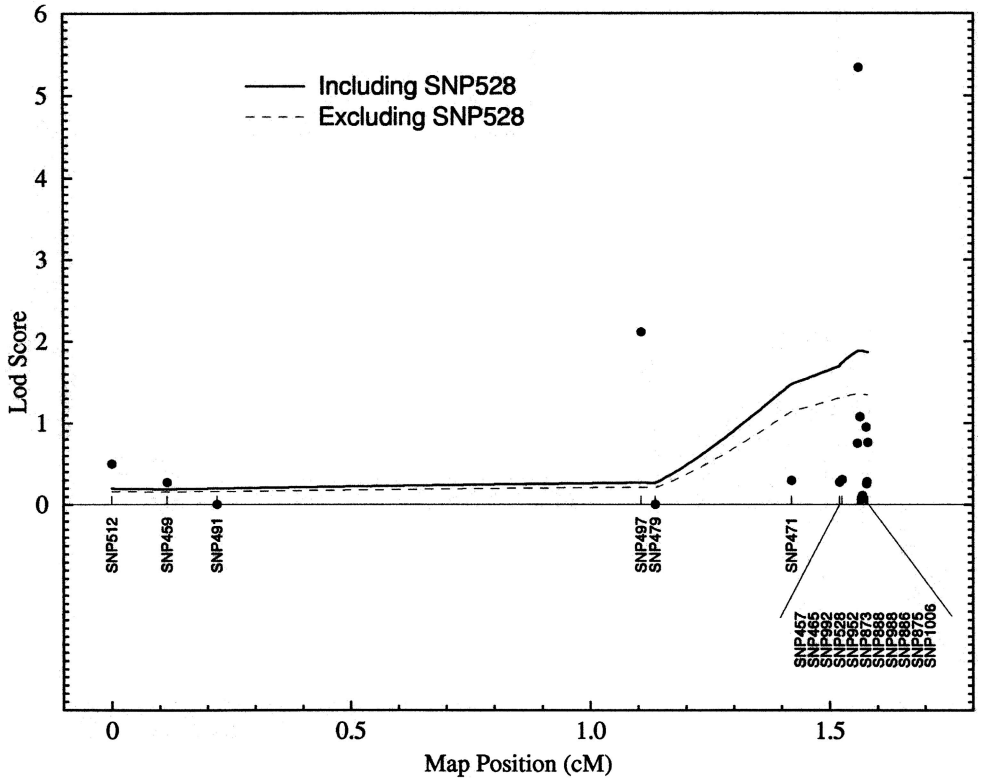

Figure 5 Two-point (*dots*) and multipoint (*lines*) LOD scores from linkage analyses

Laboratory (Murray et al. 1994) versus the integrated genetic map (Ashworth et al. 1995) shows that the approximation 1Mb = 1cM may not be unrealistic for the region surrounding *APOE*. Parametric two-point LOD scores for dominant and recessive affecteds-only models assuming a disease-allele frequency of .15 were calculated by FASTLINK (version 4.0) (Lathrop et al. 1984).

Results

Table 1 shows distances, calculated by radiation-hybrid analysis, between the SNPs and APOE-4 (SNP528), estimates of allele frequencies in controls, and *P* values both for tests of Hardy-Weinberg disequilibrium (in the cases, in the controls, and in the combined sample), and for case-control tests (based on genotypes and on alleles) for association with AD. The SNPs analyzed span ~1.5 Mb distal to *APOE* and ~190 kb proximal. More SNPs were identified on the distal side, as a result of the high efficiency of the random sequencing approach that was used to generate SNPs on that side. No attempt was made to select markers with high allele frequencies, since allele frequency was one of the parameters of interest in this study. However, by using only seven controls to screen for polymorphisms, we naturally were biased toward detection of common polymorphisms. Nevertheless, several SNPs had quite low allele frequencies. In fact, for SNP491, there was no variation observed in the controls and only 10 copies of one of the alleles were observed in the cases.

In table 1, we have indicated *P* values that are significant at the .05 level and those that are significant at the level of .0008, which is appropriate for an overall significance level of .05 when a Bonferroni correction is used to adjust for comparisons at the 60 SNPs. Notably, because of the exploratory nature of this work, we have not attempted to correct for either multiple comparisons in the different groups (i.e., cases, controls, and combined sample) or for the different types of analyses conducted. There is little evidence for deviation from HWE at most of the markers. Only tests for Hardy-Weinberg disequilibrium at SNP501 and a few markers surrounding it are significant ($P \leq .0008$) when we correct for multiple comparisons. The minor allele for SNP501 has a very low frequency (.017 in controls), and therefore the expected genotype counts for the number of homozygotes for that allele are very small. This brings into question the appropriateness of the χ^2 test for Hardy-Weinberg disequilibrium for this marker. An exact test of significance, rather than relying on the χ^2 distribution, resulted in $P = .0002$ for the test of Hardy-Weinberg disequilibrium in the cases. Although this

Table 1

Approximate Location of SNPs, Estimate of Allele Frequencies in Controls, and *P* Values for Tests of Hardy-Weinberg Disequilibrium and for Association with AD in Cases and Controls

Locus [a]	Approximate Distance to *APOE*	Allele Frequency in Controls	*P* for Hardy-Weinberg Disequilibrium, in			*P* for Case-Control Association	
			Controls	Cases	Total	Genotypic	Allelic
SNP512	1,520 kb	.15	.34	.99	.73	.0096*	.0033*
SNP501	1,470 kb	.017	.8	5.6×10^{-20}**	6.2×10^{-13}**	.1	.55
SNP459	1,405 kb	.49	.073	.00012**	.15	.00025**	.41
SNP460	1,405 kb	.0023	.97	.97	.96	1	1
SNP461	1,405 kb	.091	.00047**	.99	.025*	.048	.18
SNP489	1,335 kb	.48	.59	.73	.84	.3	.15
SNP491	1,300 kb	0	No data	.72	.8	.0012*	.0013*
SNP517	1,240 kb	.44	.074	.077	.013*	.45	.23
SNP524	1,230 kb	.036	.6	.35	.3	.07	.078
SNP522	1,220 kb	.081	.59	.83	.58	.29	.12
SNP525	1,195 kb	.0092	.89	.78	.77	.23	.23
SNP511	1,060 kb	.13	.82	.02*	.11	.024*	.063
SNP510	1,050 kb	.13	.071	.46	.049	.24	.084
SNP509	1,030 kb	.15	.014*	.55	.15	.14	.47
SNP506	995 kb	.15	.014*	.5	.17	.13	.53
SNP507	995 kb	.16	.042*	.36	.37	.13	.49
SNP502	945 kb	.14	.018*	.22	.44	.058	.95
SNP505	935 kb	.44	.82	.89	.79	.81	.52
SNP469	900 kb	.12	.84	.22	.5	.42	.37
SNP470	900 kb	.44	.52	.91	.59	.86	.69
SNP466	830 kb	.36	.088	.02*	.0038*	.37	.15
SNP467	830 kb	.47	.79	.71	.78	.034*	.01*
SNP468	830 kb	.34	.55	.69	.84	.14	.062
SNP473	630 kb	.11	.008*	.66	.041*	.11	.12
SNP481	570 kb	.19	.8	.03*	.1	.2	.21
SNP462	535 kb	.48	.48	.91	.59	.34	.17
SNP463	515 kb	.29	.0064*	.78	.042*	.077	.24
SNP458	475 kb	.3	.98	.5	.65	.87	.84
SNP497	455 kb	.37	.07	.67	.47	.004*	.0041*
SNP479	425 kb	.3	.43	.11	.56	.019*	.026*
SNP521	240 kb	.33	.22	.99	.37	.57	.52
SNP496	230 kb	.46	.8	.67	.9	.69	.48
SNP514	210 kb	.09	.3	.79	.54	.6	.51
SNP520	185 kb	.36	.25	.73	.3	.69	.54
SNP527	160 kb	.35	.31	.069	.05	.45	.3
SNP471	140 kb	.47	.58	.32	.85	.059	.035*
SNP472	140 kb	.33	.63	.38	.37	.21	.089
SNP982	<100 kb	.39	.44	.0084*	.018*	.28	.5
SNP987	<100 kb	.04	.58	.58	.43	.99	.99
SNP457	40 kb	.47	.52	.43	.96	.21	.15
SNP464	35 kb	.22	.081	.98	.21	.16	.16
SNP465	35 kb	.41	.2	.056	.039*	.014*	.0059*
SNP874	21 kb	.13	.34	.67	.32	.74	.5
SNP992	2.5 kb	.36	.12	.042*	.023*	.0099*	.0052*
SNP528 (*APOE*)	0 kb	.14	.69	.023*	.76	8.7×10^{-20}**	5.1×10^{-20}**
SNP952	2.8 kb	.33	.86	.15	.36	.026*	.014*
SNP873	6.4 kb	.36	.094	.052	.017*	.069	.037*
SNP887	8 kb	.03	.67	.66	.54	.86	.86
SNP888	8.5 kb	.46	.95	.36	.99	4.10×10^{-5}**	8.10×10^{-6}**
SNP877	14.4 kb	.02	.79	.91	.78	.21	.21
SNP988	16 kb	.22	.56	.29	.91	4.30×10^{-9}**	6.80×10^{-10}**
SNP533	40 kb	.23	.13	.72	.39	.13	.13
SNP490	60 kb	.44	.66	.23	.56	.4	.48
SNP492	80 kb	.29	.079	.31	.053	.65	.51
SNP886	<100 kb	.3	.53	.057	.074	.53	.54
SNP875	<100 kb	.32	.48	.2	.17	.74	.55
SNP1006	<100 kb	.3	.52	.68	.98	.12	.052
SNP494	100 kb	.0023	.97	.89	.9	.16	.16
SNP535	190 kb	.13	.14	.3	.76	.22	.94
SNP474	>2 Mb	.44	.34	.41	.22	.73	.45

[a] Order is distal to proximal.

* $P \leq .05$.

** $P \leq .0008$.

Table 2

***P* Values for Family-Based Tests of Association, at 17 SNPs**

	P FOR	
LOCUS	WSDT	TRANSMIT
SNP512	.1228	.8414
SNP459	.8617	.5421
SNP491	.5637	.5342
SNP497	.9984	.8583
SNP479	.0086	.0922
SNP471	.8255	.8687
SNP457	.5765	.6075
SNP465	.1520	.1103
SNP992	.0081	.0009
SNP528 (*APOE*)	8.02×10^{-6}	4.72×10^{-11}
SNP952	.0278	.00065
SNP873	.4709	.0351
SNP888	.1723	.0085
SNP988	.0031	1.41×10^{-5}
SNP886	.2364	.2124
SNP875	.3158	.2614
SNP1006	.9946	.8872

value is not as extreme as the $P = 5.6 \times 10^{-20}$ that was observed with the χ^2 test, it still meets our criteria for significance after we correct for multiple tests.

Sixteen SNPs, including the APOE-4 polymorphism, showed evidence for association ($P \leq .05$), with the case-control test for either genotype- or allele-frequency comparisons. A cluster of seven of these SNPs lies <40 kb from *APOE*. Four of the SNPs meet the criteria for significant association with AD when we adjust for multiple tests: SNP459, SNP888, SNP988, and APOE-4 itself (SNP528), with the strongest associations being found at SNP888, SNP988, and APOE-4. Markers SNP888 and SNP988 are both found on the sequence of cosmid F19374, along with APOE-4 and APOC1 (fig. 1). Previously, we had confirmed the reported association of APOC1 with AD (Martin et al. 2000). The association at the more distant marker, SNP459, still meets our criteria for significance, yet the *P* value is orders of magnitude larger—and, hence, much less significant, than the *P* values for SNP888 and SNP988. Thus, it is unlikely that this would have led us our attention away from the region immediately surrounding *APOE*. Figure 2 shows a graphic representation of the results for the case-control test for allelic association with AD (table 1), for the SNPs closest to *APOE*. In figure 2 are shown exon and intron locations of SNPs in two genes. SNP988 and SNP888 lie in *TOM40,* which codes for the mitochondrial outer-membrane protein (LocusLink accession number 10452) (Ahting et al. 1999). In the gene for this protein, neither exons nor introns showed a greater trend toward association with AD. SNP952, in the first exon (5′UTR) of *APOE,* showed evidence for association comparable to that seen at SNP992, which lies in an intron of *APOE*.

Family-based tests for association were conducted on a subset of 17 of the 60 SNPs. Criteria included markers immediately surrounding *APOE,* as well as those markers that showed significant results in the case-control analysis. The *P* values for the two family-based tests are shown in table 2. The results from the family-based tests of association are consistent with the case-control results, in showing that significant associations are concentrated at SNPs near *APOE;* however, the family-based tests are generally less significant than the case-control tests. This is likely a result of lower power for the family-based tests, because of the smaller size of the family sample relative to the case-control sample.

Figure 3 shows the results from the haplotype analyses for TRANSMIT. Six SNPs, those falling on the cosmid and genotyped in the family data, were considered in these analyses. Bars indicating significance (minus log of the *P* value) of two-locus and three-locus haplotype analyses are shown in figure 3. Two-locus and three-locus analyses examined haplotypes for each successive set of adjacent SNPs. The analysis was conducted including SNP528 (APOE-4) (fig. 3*a*) and excluding SNP528 (fig. 3*b*) from the analysis. For comparison, single-locus test results also are shown. Not surprisingly, when we included SNP528 in the analysis, the strongest single-locus result was for SNP528 (APOE-4) ($P = 4.7 \times 10^{-11}$), and the strongest results for the haplotype analyses were for haplotypes containing SNP528 (two-locus haplotype SNP952-SNP528, $P = 3.7 \times 10^{-11}$; three-locus haplotype SNP873-SNP952- SNP528, $P = 2.6 \times 10^{-12}$). When SNP528 (APOE-4) is excluded from the analysis, the most significant results are for the haplotypes containing those markers immediately proximal to SNP528 (two-locus haplotype SNP873-SNP952, $P = 2.0 \times 10^{-6}$; three-locus haplotype SNP888-SNP873-SNP952, $P = 2.4 \times 10^{-7}$). Interestingly, the results of haplotype analyses with SNP888, SNP873, and SNP952 are notably more significant than those of the single-locus tests for association at these markers. The most significant single-locus result, if SNP528 is excluded, is for a more distant marker, SNP988 ($P = 1.4 \times 10^{-5}$). Thus, when SNP528 is excluded, the haplotype analysis both increases significance and helps to better localize the disease gene, compared with the single-locus analysis. It is interesting, however, that the haplotypes with the most-significant results do not contain the *APOE* susceptibility locus but do point to a region slightly proximal to *APOE*.

Figure 4 shows the pairwise comparisons for allelic associations between SNPs that $P < .01$. Analyses were conducted among the controls (fig. 4*a*) and among the cases (fig. 4*b*). The results follow the expected distri-

bution, along a diagonal, that would be expected for allelic associations occurring between those markers close to one another; however, it can be seen that the distance over which significant association can be detected differs greatly within the region. Significant associations were identified between SNPs as distant as 1,200 kb in the controls and 1,400 kb in the cases. Generally, associations were found spanning a larger distance in the case group than in the control group. Strong associations were identified between SNP528 (APOE-4) and both SNP988 and SNP888, in both case and control samples. This explains why, in our case-control analysis, we detected such strong evidence of association between these SNPs and AD.

Figure 5 shows the results from our linkage analyses. Two-point maximum LOD scores are shown, and two multipoint curves, one including and one excluding the APOE-4 SNP (SNP528), are shown in the plot. Clearly, the linkage analyses give good evidence for linkage to a disease gene in the region. The two-point LOD score of 5.5 at SNP528 gives very strong evidence for linkage, and the multipoint curve including SNP528 maximizes, with a LOD score of almost 2, at the position corresponding to the location of SNP528. Even when the disease polymorphism, SNP528, is excluded, there is evidence for linkage, on the basis of the multipoint analyses (LOD score >1), and the curve maximizes at the correct position. It is of note that the interval examined was not broad enough to allow us to bound the linkage on the proximal side of *APOE*. Results from parametric linkage analyses (data not shown) were qualitatively similar to those from nonparametric linkage analyses. Both dominant and recessive models give the strongest two-point LOD scores for SNP528.

Discussion

This study demonstrates that associations can be detected at SNPs near a complex disease gene. However, it is clear that, in order to detect association, the "right" SNPs must be chosen. Of 13 SNPs genotyped at <40 kb from the APOE-4 polymorphism, 7 (including APOE-4) showed evidence of association with AD ($P \leqslant .05$). The APOE-4 SNP and two other SNPs lying <16 kb from APOE-4 showed strong evidence of association. Since the strongest associations were detected <16 kb from the disease polymorphism, does this suggest that a map density of 32 kb would be adequate for detection of association with AD? Several of the SNPs closest to the APOE-4 SNP showed only marginal evidence or no evidence of association with AD. Thus, even assuring that one SNP <16 kb of the disease locus was chosen would not guarantee that an association could be identified. Using a greater density of markers would be desirable, since it increases the chance of including any markers with significant association with the disease. For this reason, the efforts underway to identify hundreds of thousands of SNPs spanning the genome appear to be well justified. Of course, it should be noted that this is a very limited example, and the required density of markers will surely vary depending on the disease under study and the region of the genome being analyzed.

It has been suggested that, for identification of complex disease loci, SNPs in coding regions may be more useful than those in noncoding regions. However, we find that, unless a functional SNP directly influencing disease susceptibility (e.g., APOE-4) is included in the battery of markers tested, there is no clear advantage to SNPs in exons rather than in introns. Thus, this analysis strongly suggests that success in finding an association rests on having a high density of SNPs. As can be seen in figure 4, areas of association will differ along the chromosome, and this may be useful in predicting the necessary density of SNPs for studies.

The reduction in the number of markers demonstrating significance after correction for multiple comparisons is an important factor to consider. Although the results of the tests for association at SNP988 and SNP888 are clearly significant in the case-control analyses, it is easy to see that, if one is performing analysis of a larger number of SNPs, the noise level is a significant problem. In fact, it is likely that, in a genomic screen using as few as 10,000 SNPs, the test for association at SNP888 would not be significant if a standard Bonferroni correction for multiple comparisons were applied. It is important, however, to consider the tests for association at the multiple markers considered jointly. Several SNPs immediately surrounding *APOE* show marginal significance, with $P \leqslant .05$. Perhaps seeing a region of markers with P values nearing significance for association is an important indicator of the presence of a disease gene and, as has been suggested for linkage analysis, will help distinguish true positives from false positives (Terwilliger et al. 1997).

Haplotype analysis to investigate associations between disease loci and multiple markers could be a valuable tool for the use of SNPs in complex disease. In this study, in the families with AD, the haplotype analyses conducted for SNPs around *APOE* did increase significance and improve localization of the disease gene. This is evidenced by the fact that analyses of two- and three-locus haplotypes yielded results that were more significant than those of the single-locus tests. Furthermore, localization, when APOE-4 is not included in the analysis, is improved by use of haplotypes, compared with the use of single loci. These results suggest that association analysis based on haplotypes may indeed be a powerful approach for mapping of complex disease genes by use of SNPs.

The results for the family-based tests of association

were generally less significant than the results for the case-control test. This was not unexpected, since there are a larger number of independent observations in the case-control sample. In general, larger samples will need to be obtained for family-based tests of association (Martin et al. 1997). However, there are advantages to a family-based design, such as robustness in stratified populations, utility for both linkage and association analyses, and the ability to conduct haplotype analyses, which may outweigh the additional cost of the collection of larger samples.

It is difficult to say much about the effect of SNP allele frequencies on association analysis; however, it does appear that it may be preferable to use SNPs with moderate allele frequencies, rather than SNPs with very low allele frequencies. For example, in the case-control analysis, SNP988 shows highly significant association with AD, yet neither SNP877 nor SNP887, both of which are closer to APOE-4, shows evidence of association with the disease. This could be due to their low allelic frequencies, both of which are estimated to be <.05 in controls (table 1). In fact, in the case-control test, of the eight genotyped SNPs <20 kb from APOE-4, only SNP877 and SNP887 have nonsignificant results, whereas the other six SNPs, all with allele frequencies >.14 in controls, show evidence of association ($P \leq .05$). Allele frequencies also influence the power of the family-based tests for association. Sibships in which all siblings have the same marker genotype are not informative. Thus, there will be fewer informative families for markers with lower heterozygosity; for example, for SNP491, only four informative families were observed, and it is not surprising that the results of family-based tests for association at this marker were not significant. These data suggest that markers with very low allele frequencies may not be particularly useful for association analysis. However, this is not to suggest that only SNPs with allele frequencies of ~.5 should be chosen. SNP528 and SNP988 have estimated allele frequencies of .14 and .22, respectively, in the controls; however, the case-control and family-based tests for association at these SNPs are more significant than the tests for all of the other SNPs, many of which are more polymorphic.

One very interesting finding is that SNP952 and SNP992, two SNPs in the *APOE* gene itself, are not highly associated with AD. SNP952 is in the 5′UTR, and SNP992 is intronic, suggesting that it is unlikely that any selective pressures would maintain one allele rather than another; although low allele frequency would appear to be an obvious explanation of this finding, the minor allele frequencies for these two SNPs were .36 and .33, respectively (table 1). Thus, low allele frequency does not appear to be a factor in this finding. This example does suggest, however, that ruling out a candidate gene as contributing to a disease state, particularly when SNPs in noncoding regions are used, may be very difficult.

SNP988 (coding) and SNP888 (noncoding), two SNPs showing strong evidence of association with AD, lie in the *TOM40* gene. The protein for which this gene codes constitutes the primary component of the protein-conducting channel of the outer mitochondrial membrane. However, the SNP988 base change is silent, producing no change in the amino acid coded. Thus, as in the case of APOC1, it is likely that the association with AD is the result of linkage disequilibrium with the APOE-4 allele. This illustrates the difficulty that can be present in association studies when positive associations with multiple biologically relevant genes are found.

All of the analyses conducted in this study give strong evidence for existence of an AD gene in the region. The results of association analyses would have allowed us to narrow the region of interest to a relatively small region containing the *APOE* gene. However, it is important to keep in mind that, even with the region narrowed, there is still the challenge of identifying the actual gene involved. Findings of association with *APOE* and AD have been widely replicated, and several lines of evidence support a functional role of the *APOE* gene in AD (Schmechel et al. 1993; Strittmatter et al. 1993*a*, 1993*b*). It was the convergence of positional and biological evidence, followed by the demonstration of association with the functional polymorphism itself, that ultimately led to the identification of the role of *APOE* in AD. Therefore, a multidisciplinary approach for identification of genes influencing complex disease will most useful.

Our analysis of SNPs in this region has clearly demonstrated that it is possible to detect association at markers near a disease polymorphism, even for a complex disease in outbred populations. This study demonstrates the need for a high density of SNP markers and has found a benefit in the use of markers jointly in haplotype analyses. We are continuing this investigation by using this model, investigating the effects of haplotypes and linkage disequilibrium in the region for AD, in specialized populations.

Acknowledgments

We thank all of the families whose participation made this project possible. This research was supported in part by National Institutes of Health (NIH) Program Project grant 2 P01 NS26630-11A1, NIH/National Institute of Neurological Disorders and Stroke grant 5 R01 NS31153-07, National Institute on Aging grant 5 P50 AG05128-16, and funding from Glaxo Wellcome, Inc. Support also was provided by grants from the Alzheimer's Association. We also thank the personnel at the Center for Human Genetics of Duke University Medical Cen-

ter, and special thanks are due to Drs. Beth Hauser, Bill Scott, Dmitri Zaykin, and Norman Kaplan, for the many discussions regarding this work.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html (for cosmid sequences)
GDA: Software for the Analysis of Discrete Genetic Data, http://alleyn.eeb.uconn.edu/gda
LocusLink, http://www.ncbi.nlm.nih.gov/LocusLink (for *TOM40* [accession number 10452])
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim (for AD [MIM 104300] and *APOE* [MIM 107741])

References

Ahting U, Thun C, Hegerl R, Typke D, Nargang FE, Neupert W, Nussberger S (1999) The TOM core complex: the general protein import pore of the outer membrane of mitochondria. J Cell Biol 147:959–968

Ashworth LK, Batzer MA, Brandriff B, Branscomb E, De Jong P, Garcia E, Garnes JA, et al (1995) An integrated metric physical map of human chromosome 19. Nat Genet 11: 422–427

Blacker D, Haines JL, Rodes L, Terwedow H, Go RCP, Harrell LE, Perry RT, et al (1997) ApoE-4 and age at onset of Alzheimer's disease: the NIMH genetics initiative. Neurology 48:139–147

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229–1231

Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC Jr, Rimmler JB, et al (1994) Protective effect of apolipoprotein e type 2 allele for late onset Alzheimer disease. Nat Genet 7:180–184

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Freitas EM, Zhang WJ, Lalonde JP, Tay GK, Gaudieri S, Ashworth LK, Van Bockxmeer FM, et al (1998) Sequencing of 42kb of the APO E-C2 gene cluster reveals a new gene: PEREC1. DNA Seq 9:89–101

Gerdes LU, Gerdes C, Hansen PS, Klausen IC, Færgeman O, Dyerberg J (1996) The apolipoprotein E polymorphism in Greenland Inuit in its global perspective. Hum Genet 98: 546–550

Hanlon CS, Rubinsztein DC (1995) Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. Atherosclerosis 112:85–90

Hauser ER, Boehnke M (1998) Genetic linkage analysis of complex genetic traits by using affected sibling pairs. Biometrics 54:1238–1246

Heid CA, Stevens J, Livak KJ, Williams R (1996) Real time quantitative PCR. Genome Res 6:986–994

Lai E, Riley J, Purvis I, Roses A (1998) A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics 54:31–38

Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

Marshall E (1999) Genomics: drug firms to create public database of genetic mutations. Science 284:406–407

Martin ER, Gilbert JR, Lai EH, Riley J, Slotterbeck BD, Rogala AR, Sipe CA, et al (2000) Analysis of association at SNPs in the APOE region. Genomics 63:7–12

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61: 439–448

Martin ER, Monks SA, Warren LL, Kaplan NL (1999) A weighted sibship disequilibrium test for linkage and association in discordant sibships. Am J Hum Genet Suppl 65: A434

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddma TS, Manion F, Quillen J, et al (1994) A comprehensive human linkage map with centimorgan density: Cooperative Human Linkage Center (CHLC). Science 265: 2049–2054

Pericak-Vance MA, Bass MP, Yamaoka LH, Gaskell PC, Scott WK, Terwedow HA, Menold MM, et al (1997) Complete genomic screen in late-onset familial Alzheimer disease: evidence for a new locus on chromosome 12. JAMA 278: 1237–1241

Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, et al (1991) Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. Am J Hum Genet 48:1034–1050

Rimmler J, McDowell JG, Slotterback BD, Haynes CS, Menold MM, Rogala A, Speer MC, et al (1998) Development of a data coordinating center (DCC): data quality control for complex disease studies. Am J Hum Genet Suppl 63:A240

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

Saunders AM, Schmader KE, Breitner JC, Benson MD, Brown WT, Goldfarb L, Goldgaber D, et al (1993) Apolipoprotein E ϵ4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. Lancet 342:710–711

Schmechel DE, Saunders AM, Strittmatter WJ, Crain BJ, Hulette CM, Joo SH, Pericak-Vance MA, et al (1993) Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late onset Alzheimer disease. Proc Natl Acad Sci USA 90:9649–9653

Strittmatter WJ, Saunders AM, Pericak-Vance MA, Salvesen GS, English J, Roses AD (1993*a*) Apolipoprotein E: high avidity binding to a bA amyloid and increased frequency of type 4 isoform in familial Alzheimer's disease. Proc Natl Acad Sci USA 90:1977–1981

Strittmatter WJ, Weisgraber KH, Huang DY, Dong LM, Salvesen GS, Pericak-Vance MA, Schmechel D, et al (1993*b*) Binding of human apolipoprotein E to synthetic amyloid b peptide: isoform-specific effects and implications for late-

onset Alzheimer disease. Proc Natl Acad Sci USA 90: 8098–8102

Terwilliger JD, Shannon WE, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE (1997) True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. Am J Hum Genet 61:430–438

Vance JM, Martin ER, Lai E, Riley J, Slotterbeck BD, Sipe CA, Barker JM, et al (1998) Genotyping and association studies of single nucleotide polymorphisms (SNPs) in a 4 megabase region surrounding the Alzheimer's disease (AD) risk factor, APOE. Am J Hum Genet Suppl 63:A312

Zaykin D, Zhivotovsky L, Weir BS (1995) Exact tests for association between alleles at arbitrary numbers of loci. Genetica 96:169–178