Am. J. Hum. Genet. 67:1437–1451, 2000

Haplotypes at *ATM* Identify Coding-Sequence Variation and Indicate a Region of Extensive Linkage Disequilibrium

Penelope E. Bonnen,[1] Michael D. Story,[2] Cheryl L. Ashorn,[2] Thomas A. Buchholz,[3] Michael M. Weil,[2] and David L. Nelson[1]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, and Departments of [2]Experimental Radiation Oncology and [3]Radiation Oncology, MD Anderson Cancer Center, Houston

Genetic variation in the human population may lead to functional variants of genes that contribute to risk for common chronic diseases such as cancer. In an effort to detect such possible predisposing variants, we constructed haplotypes for a candidate gene and tested their efficacy in association studies. We developed haplotypes consisting of 14 biallelic neutral-sequence variants that span 142 kb of the *ATM* locus. *ATM* is the gene responsible for the autosomal recessive disease ataxia-telangiectasia (AT). These *ATM* noncoding single-nucleotide polymorphisms (SNPs) were genotyped in nine CEPH families (89 individuals) and in 260 DNA samples from four different ethnic origins. Analysis of these data with an expectation-maximization algorithm revealed 22 haplotypes at this locus, with three major haplotypes having frequencies $\geqslant$.10. Tests for recombination and linkage disequilibrium (LD) show reduced recombination and extensive LD at the *ATM* locus, in all four ethnic groups studied. The most striking example was found in the study population of European ancestry, in which no evidence for recombination could be discerned. The potential of *ATM* haplotypes for detection of genetic variants through association studies was tested by analysis of 84 individuals carrying one of three *ATM* coding SNPs. Each coding SNP was detected by association with an *ATM* haplotype. We demonstrate that association studies with haplotypes for candidate genes have significant potential for the detection of genetic backgrounds that contribute to disease.

Introduction

Qualifying and quantifying the genetic contribution to the etiology of common complex disease remains one of the great quests of modern medical genetics. The complexity of multifactorial diseases challenges the paradigms and tools of conventional genetic research. Traditional methods of genetic analysis do not have the statistical power or sensitivity for the task of teasing out a genetic contribution when it is subtle or when several genes may be working together (Risch and Merikangas 1996). Genomewide association studies, as well as population studies with candidate genes, have been touted as possible alternatives to linkage analysis (Risch and Merikangas 1996; Collins et al. 1997; Kruglyak 1999; Risch 2000). These approaches focus on finding either a causative variant or a genetic variant closely linked with the disease phenotype. Some studies utilizing single-nucleotide polymorphisms (SNPs) have succeeded in detecting the risk for disease, notably in the case of the apolipoprotein type E (apoE) gene and both coronary artery disease (Boerwinkle et al. 1996) and Alzheimer disease (Strittmatter and Roses 1995). These studies were able to directly assess the risk conferred by known apoE functional variants. In some other cases, however, the attempt to correlate single-locus alleles with phenotypes have produced mixed results (Josefsson et al. 1998; Kraft et al. 1998; Storey et al. 1998).

Haplotype association with disease by the linkage disequilibrium (LD) approach has been used successfully for the identification of genomic regions containing loci responsible for disease phenotypes (MacDonald et al. 1992; Yu et al. 1996). The same principle can be applied by use of haplotypes of biallelic markers to detect disease association. Using several SNPs distributed across 100–200 kb should result in statistical sensitivity that is greater than that in studies using fewer loci. Another strength of such an approach is the ability to use purely epidemiological populations for detection of chromosomal backgrounds lending risk for disease.

All of these approaches are, to one extent or another, dependent on LD. An understanding of LD relationships between markers will inform the efficacy and design of future LD-based strategies for detection of genetic contributions to common disease. Simulation studies have estimated the length of useful LD to be as low as 3 kb (Kruglyak 1999). Recent investigations support the no-

Received August 22, 2000; accepted for publication September 21, 2000; electronically published November 14, 2000.

Address for correspondence and reprints: Dr. David L. Nelson, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030. E-mail: nelson@bcm.tmc.edu

0002-9297/2000/6706-0010$02.00

tion that LD varies throughout the genome (Collins et al. 1999; Taillon-Miller et al. 2000) and that it can extend to considerable lengths, such as several hundred kilobases (Collins et al. 1999; Eaves et al. 2000; Moffatt et al. 2000; Taillon-Miller et al. 2000). Reports of such extreme differences indicate the need for further study of the extent and nature of LD.

Allelic variation leading to functional variants of genes may predispose to risk for seemingly sporadic cases of common disease (Lander 1996; Collins et al. 1997). Here we describe a strategy for exploring the possible effects of functional variants of genes involved in familial cancers. We use a resequencing approach to detect SNPs across a large (184 kb) genomic region containing the *ATM* gene. *ATM* is responsible for the autosomal recessive disease ataxia-telangiectasia (A-T) (Savitsky et al. 1995). A-T is characterized by cerebellar ataxia, oculocutaneous telangiectasia, immune deficiency, sensitivity to ionizing radiation, increased incidence of tumors, and chromosomal instability (Gatti et al. 1991). A-T heterozygotes may be at increased risk for development of cancers, most prominently—and controversially—breast cancer (Swift et al. 1987, 1991; Morrell et al. 1990; Stankovic et al. 1998; Gatti et al. 1999). With carrier frequencies estimated to be from 0.5% to >1% (Swift et al. 1986; Gatti et al. 1999), assessment of cancer risk for this population is a compelling endeavor. In addition, the *ATM*-gene product is centrally involved in cellular responses to DNA damage, including DNA double-strand break repair and signaling leading to cell-cycle arrest and apoptosis (reviewed in Rotman and Shiloh 1999). We genotyped 295 individuals from four ethnic groups, for 14 SNP markers that spanned 142 kb. An expectation-maximization algorithm estimated 22 *ATM* haplotypes from these data. Tests for recombination and LD revealed (*a*) no evidence for recombination in the white European American study population and (*b*) perfect disequilibrium extending the full length marked by these SNPs. We then conducted a model association study with these haplotypes and a population of samples that possessed one of three different coding SNPs (cSNPs) in the *ATM* gene. The results of this study provide strong support for the utility of complex SNP haplotypes as a means to detect polymorphisms in a population-based sample.

Subjects and Methods

Human Subjects

For SNP discovery, genomic DNA from five unrelated white European Americans was sequenced. This DNA was extracted from lymphoblast and fibroblast cell lines. For SNP genotyping, individuals from four ethnic groups were sampled: African American ($n = 71$), Asian American ($n = 39$), white European American ($n = 77$), and Hispanic ($n = 73$). All ethnic samples (self-described ethnicity) were part of a collection of 941 DNA purified samples from anonymous blood donors in community-based blood drives in southeastern and central Texas. Samples analyzed in the model association study were also from this DNA collection. Members of nine CEPH families were also analyzed. In all families, four grandparents, two parents, and four children were examined; since two of these families share a grandparent, 89 individuals were genotyped, and the number of segregating chromosomes is 70.

Samples from Great Apes

Six great-ape samples were genotyped in this study: two from common chimpanzees (*Pan troglodyte*), one from a bonobo (*P. paniscus*), two from western lowland gorillas (*Gorilla gorilla*), and one from an eastern lowland gorilla (*G. g. graueri*).

PCR and Sequencing Primers

Primers for DNA amplification and sequencing were designed by MacVector, version 6.0.1. The 184-kb genomic sequence of *ATM* was masked for repetitive sequence, by Repeat Masker. Thirty-six primer sets were designed to amplify regions containing little or no repeat sequence, distributed evenly throughout the sequence. Primers were selected that met strict criteria for melting temperature and that amplified regions containing very little or no repeat sequence. The same primers were used for PCR and sequencing reactions and are listed in Appendix A.

PCR Amplification of Genomic DNA

Genomic DNA from five unrelated individuals was amplified by means of 29 of the 36 primer sets mentioned above. The 50-μl reactions included DNA (200 ng), standard PCR buffer, dNTPs (0.1 mM each), *Taq* (0.5 μl; Perkin-Elmer), and primers (1 μM each). PCR was performed in a Perkin Elmer 9700 analyzer, with an initial denaturation at 95°C for 5 min, followed by 30 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s, and a final step at 72°C for 7 min. For all amplicons, 6 μl of PCR product was run on a 1.5% agarose gel.

DNA Sequencing

PCR products were purified and sequenced. Preparation of DNA for sequencing included incubation of ~60 ng of PCR product with shrimp alkaline phospha-

tase (2 U; Amersham) and exonuclease I (10 U; Amersham) at 37°C for 15 min, followed by enzymatic inactivation at 80°C for 15 min. Sequencing of each PCR product was performed with the Thermo Sequenase [^{33}P]-radiolabeled terminator-cycle sequencing kit (Amersham Pharmacia), according to the manufacturer's instructions. Sequencing reactions were performed in a Perkin Elmer 9700 analyzer, with an initial denaturation at 95°C for 1 min, followed by 35 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min. Samples were run on 6% polyacrylamide gels, fixed for 15 min in 5% acetic acid/20% methanol, and dried.

Multiplex PCR

Sequencing revealed 17 SNPs in 15 different regions of the gene. These 15 PCR amplicons were multiplexed into two PCR reactions. Multiplex group 8 amplifies eight fragments, and Multiplex group 7 amplifies seven fragments. The 50-μl reactions for group 7 included DNA (400 ng), standard PCR buffer (2 ×), dNTPs (0.2 mM each), and *Taq* (0.5 μl; Perkin-Elmer). The 50-μl reactions for group 8 included DNA (400 ng), standard PCR buffer (1.8 ×), dNTPs (0.2 mM each), and *Taq* (0.5 μl; Perkin-Elmer). Primers include some of those originally designed for sequencing and some of those newly designed to alter the size of the amplicons. Products were separated by $\geqslant$20 bp, so that they could be resolved from one another on a 2.5% agarose gel. Multiplex PCRs were checked to have amplified all products, by running 6 μl of product on a 2.5% agarose gel. The concentrations and primer sequences used for PCR are listed in Appendix B.

Allele-Specific Oligonucleotide (ASO) Hybridizations

Genotypes for each SNP were determined in all sample populations, by ASO hybridizations. ASO hybridizations were performed as described by DeMarchi et al. (1994). We performed ASO hybridization for 14 SNPs for each individual typed. These 14 SNPs were chosen from the original 17 because they perform consistently well under standard ASO-hybridization conditions. Hybridizations were performed under conditions that allowed for annealing of only the probe that is an exact match for the substrate DNA. Genotypes for SNPs were read on at least two independent occasions. The sequences of the ASO-hybridization probes are listed in Appendix C.

Estimation of Haplotypes and Frequencies

Haplotypes and their frequencies were estimated on the basis of unphased genotype data, by the computer program EMHAPFRE. Described in the work of Excoffier and Slatkin (1995), EMHAPFRE uses an expectation-maximization algorithm that determines the maximum-likelihood frequencies of multilocus haplotypes in diploid populations. Only individuals who were scored for all 14 SNPs were included in the data analysis.

Haplotype Assignment to Genotype Data

A short script written in Microsoft Excel Visual Basic and named "Assign" was used to assign genotypes to individual samples. The script was given, as input, the list of haplotypes produced by EMHAPFRE and the raw unphased genotype data. It produces a list of samples input, with a pair(s) of haplotypes that satisfies the genotype data assigned to each; in cases in which multiple pairs of haplotypes were listed, one pair is chosen, by use of a haplotype frequency–based method. A probability is calculated for each haplotype pair, by multiplication of the haplotypes' frequencies in the control population. The haplotype pair with the highest probability is assigned to the individual.

Statistical Analysis for Recombination and LD

To test for recombination, we used the four-gamete test and the Hudson and Kaplan (1985) recombination statistic, R. For a given haplotype AB, mutation may result in either Ab or aB. Haplotype ab arises only in the case of either recombination or repeat mutation. The four-gamete test was executed on unphased genotype data, in a pairwise fashion, across all SNP loci. On the basis of the resulting matrix of the four-gamete test, R estimates the location and number of recombination events that have occurred in the sample.

Initial LD analysis was computed by performance of pairwise comparisons for all SNP loci. Fisher's exact test was used to determine significance levels. SNPs having a minor-allele frequency of .05 were excluded from LD analyses. LD statistic D is a pairwise comparison of gametic frequencies such that D = p11p22–p12p21. D', the relative disequilibrium, is $D' = D/|D|\max$, where $|D|\max = \max(\text{p1p2,q1q2})$ if $D < 0$ and $|D|\max = \min(\text{q1p2,p1q2})$ if $D > 0$. D' ranges from 1 to -1, and this range is not influenced by allele frequency.

All recombination and LD statistics were generated by the software program DnaSP 3.00 (written by J. Rozas and R. Rozas, University of Barcelona).

Statistical Analysis for Association Study

Testing for significance in the model association study was done by use of contingency tables for independence. P values for significance of association at the haplotype level were determined by use of 2 × 2 tables and 3 × 3 tables for the genotype level. Significance values refer to a one-sided test.

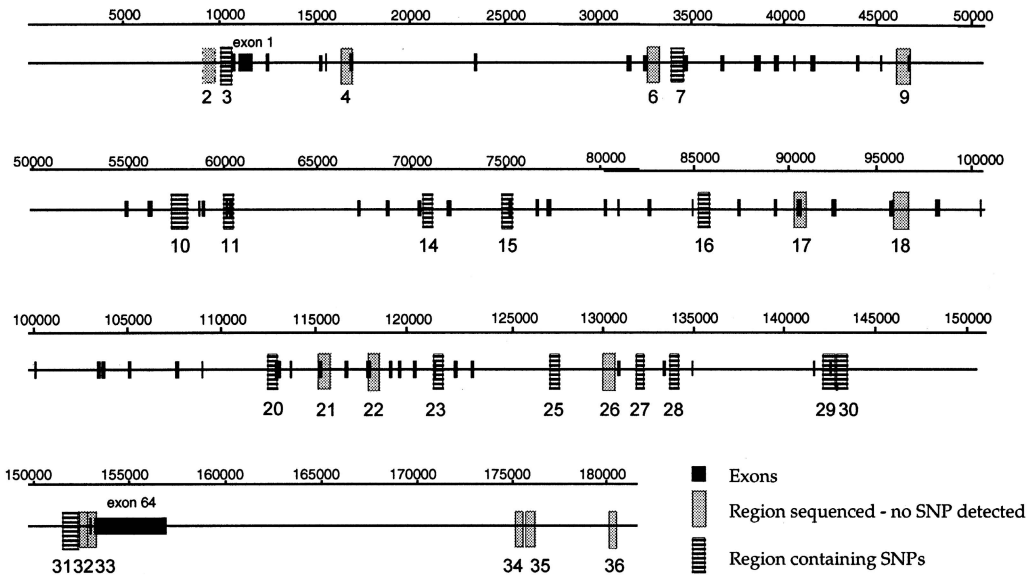

Figure 1 Schematic of *ATM*. The 184 kb of the *ATM* locus is illustrated, with the 64 exons represented by black boxes. Twenty-nine ~500-bp regions were amplified by PCR in five unrelated individuals. These regions were sequenced and found to contain 17 SNPs.

Results

SNP Discovery

Our initial objective was to discover common neutral sequence variants spanning the length of the *ATM* gene. A gel-based resequencing strategy was employed to detect SNPs at the *ATM* locus. Genomic DNA of five unrelated individuals was amplified, by PCR, for [^{33}P]-radiolabeled sequencing. For detection of markers spanning the entire locus, PCR primers were designed for amplicons dispersed approximately evenly throughout the 184-kb genomic region containing the gene (fig. 1). Approximately 13.5 kb of the 184-kb total sequence was read in each individual. The nucleotide diversity, π, calculated for this sequence data was .00057. Seventeen SNPs were found, which span 142 kb and all of which are located in introns (table 1). This yielded an average of 1 SNP/794 nucleotides sequenced.

Genotyping and Haplotype Development

To begin construction of haplotypes from these SNPs, we genotyped nine three-generation CEPH families (Dausset et al. 1990). By using three-generation families, we could determine haplotypes from genotype data, through inference. This allowed us both to determine the efficacy of the computer algorithm used to predict haplotypes (see below) and to optimize our genotyping assay. We performed ASO hybridization on nine CEPH families (89 individuals; 70 chromosomes), for 14 of the original 17 SNPs. These 14 SNPs were chosen from the original 17 because they performed consistently well under standard ASO-hybridization conditions.

We then used two different methods for deciphering the haplotypes derived from the genotype data, in a side-by-side comparison. First, haplotypes were inferred by

Table 1

Seventeen *ATM* Noncoding SNPs Detected by Resequencing

SNP[a]	Location in Genomic Sequence with GenBank Accession Number U82828
Prior to 5′UTR t→a[b]	10182
IVS8−356t→c	34293
IVS19−1276a→g	57469
IVS21−77t→c	60136
IVS26+491c→g[c]	71049
IVS27−193c→t[c]	75083
IVS34+754g→a	85811
IVS46−257a→c	112721
IVS55+186c→t	121819
IVS57+3570t→c	127195
IVS58+997g→a	132032
IVS59+414g→t[c]	133986
IVS61−55t→c	142611
IVS62+60g→a	142789
IVS62+424g→a	143153
IVS62−973a→c	151964
IVS62−694c→a	152243

[a] Nomenclature is according to the guidelines recorded by the Ad Hoc Committee on Mutation Nomenclature (1996).

[b] This SNP is named in reference to the genomic sequence having GenBank accession number U82828 because of the highly variable nature of the 5′UTR.

[c] Not used in genotyping or haplotype analysis.

Table 2

***ATM* Haplotypes of 295 Humans from Five Ethnic Groups and of Three Species of Great Apes**

		Frequency in Humans[a]					
Haplotype	Sequence	Overall (n = 295)	African American (n = 71)	Asian American (n = 39)	White European American (n = 77)	Hispanic American (n = 73)	CEPH (n = 35)
1	ACTCTACTTCCCTC	.002				.007	
2[b]	ACTCTACTTCTTTC	.313	.190	.500	.292	.315	.394
3	ACTCTCCTTCTTTC	.037			.065	.048	.061
4	ACTTCACTCCTCTC	.002	.007				
5	ACTTTACTCTCCTC[c]	.002					
6[b]	ACTTTACTTCCCTC	.066	.218	.013	.013	.027	.015
7	ACTTTACTTCTTTC	.019	.077				
8	ATTCTACTTCTTTC	.012	.007	.051		.007	
9	ATTCTCCTTCTTTC	.000		.013			
10	ATTTCACTCCCCTC	.002	.007				
11	ATTTCATCCTCCCC	.002		.013			
12	TCTCTACTTCTTTC	.007				.021	.015
13	TCTTCACTCTCCTC	.010	.035			.007	
14	TCTTCATCCTCCCC	.002				.007	
15[b]	TTCTCACTCTCCTA	.090	.028	.013	.175	.041	.227
16	TTTCTATCCTCCCC	.005		.017		.007	
17[b]	TTTTCACCCTCCTC	.100	.141	.068	.097	.110	.015
18	TTTTCACCCTCTTC	.002				.007	
19	TTTTCACTCCTTTC	.002	.007				
20	TTTTCACTCTCCTA	.002				.007	
21[b]	TTTTCACTCTCCTC	.048	.162	.013	.006	.027	
22[b]	TTTTCATCCTCCCC	.277	.113	.291	.351	.363	.273
	TTTCTACCCTCCTC[c]	…		.009			
	ACTTTACCCTCCTC[c]	…	.007				
Total		1.000	1.000	1.000	1.000	1.000	1.000

		Frequency in Great Apes[d]		
		Chimpanzee (n = 2)	Bonobo (n = 1)	Gorilla (n = 3)
1	TCTTTACTCTCCTC	.750	1.000	.000
2	TCTTTACTCTCTTC	.250	.000	.000
3	TATTTACTCTCCTC	.000	.000	1.000

[a] Samples were genotyped by ASO hybridization, then haplotypes and their frequencies were estimated from unphased genotype data, by the EM algorithm EMHAPFRE.

[b] Haplotype present in all four ethnic groups studied.

[c] Low-frequency haplotypes in which some differences were seen in the combined data set and in individual ethnic populations.

[d] Samples were genotyped by ASO hybridization and fluorescent sequencing.

hand. We began with homozygotes and predicted other haplotypes on the basis of transmission and by establishing the phase through the pedigrees. Seven haplotypes were identified in the sample of CEPH families. Subsequently, we subjected the same data set to an expectation-maximization algorithm, to estimate haplotypes and their frequencies. The computer program EMHAPFRE is a maximum-likelihood program developed to predict multilocus haplotypes from unphased genotype data (Excoffier and Slatkin 1995). It produces both a list of haplotypes and their estimated frequencies in the input sample population. The haplotype predictions from EMHAPFRE were in complete accordance with those that had been inferred manually, giving us confidence that this program was suitable for data of this nature.

Haplotype and Allele Frequencies

To determine frequencies of haplotypes and of individual SNPs in different ethnic populations, we performed ASO hybridization on anonymous African American (n = 71), Asian American (n = 39), white European American (n = 77), and Hispanic (n = 73) DNA samples collected in central and southeastern Texas. Genotype data were analyzed by the EMHAPFRE program. For the total population, 22 haplotypes and their frequencies were predicted by EMHAPFRE (table

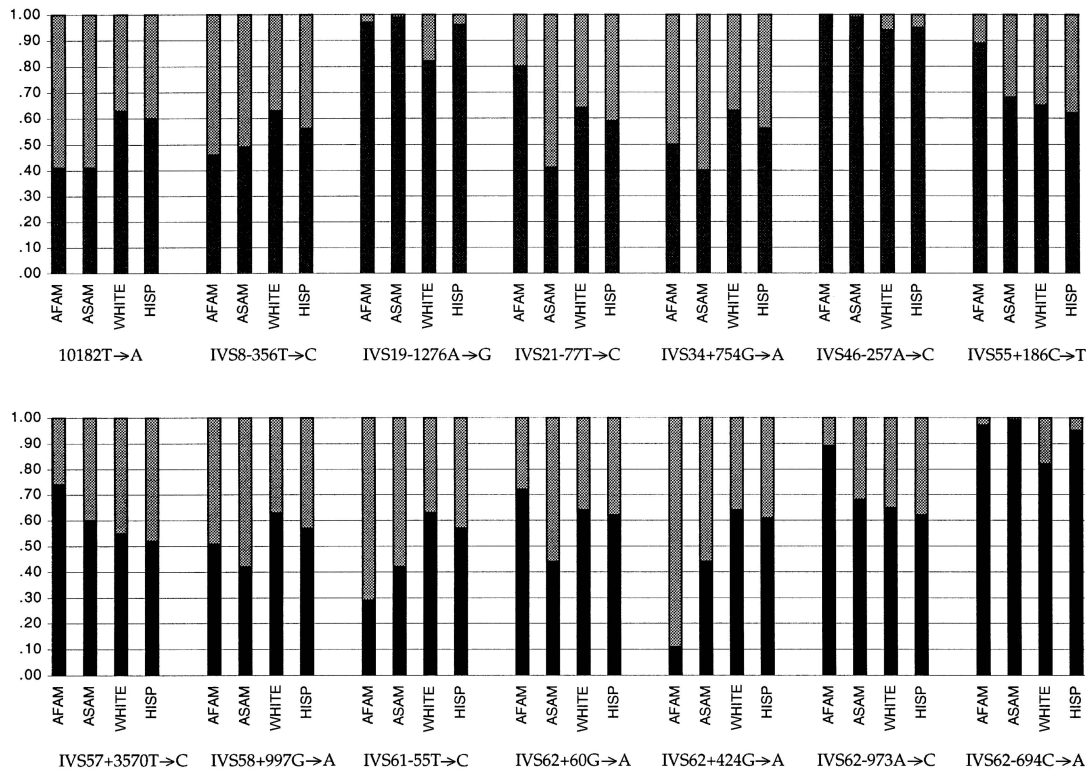

Figure 2 ATM SNP allele frequencies for 14 ATM SNPs in each of four ethnic groups. A total of 260 individuals (71 African American, 39 Asian American, 77 white European American, and 73 Hispanic) were genotyped by ASO hybridization.

2). Three predominant haplotypes were found at frequencies ⩾10%. An independent study that examined neutral sequence variants at the *ATM* locus also found three major haplotypes (Li et al. 1999).

The majority of SNPs identified in this study have a frequency, in all ethnic groups, of ⩾25% (fig. 2). Of the 14 SNPs, 3 (IVS19−1276a→g, IVS46−257a→c, and IVS62−694c→a) have a minor-allele frequency of <10% in most ethnic groups. SNP frequencies vary across ethnic groups. Three SNPs (IVS55+186c→t, IVS62+424g→a, and IVS62−973a→c) have a frequency of 11% in African Americans while being present at a frequency of >30% in all other ethnic groups. SNP IVS46−257a→c was not found in the samples from African Americans. Of the three low-frequency SNPs, two (IVS19−1276a→g and IVS62−694c→a) have a frequency of >18% in the white European American population and of <6% in the others. This is not surprising, given that the original five samples used for SNP detection were white European Americans.

To begin to describe the haplotype phylogeny at the *ATM* locus, we wanted to determine what haplotypes were present in each ethnic population. The genotype data were analyzed, by EMHAPFRE, as four separate data sets segregated by ethnic group. However, this analysis led to small discrepancies from what was predicted from the complete data set. In each case, changes were found in the lowest-frequency haplotypes (table 2). The efficacy of EMHAPFRE is known to decay as data sets decrease in size (Excoffier and Slatkin 1995). Thus, a second approach to ascription of haplotypes and their frequencies to each ethnic group was taken. To this end, a simple script was written in Microsoft Excel Visual Basic. This script, named "Assign," takes a list of haplotypes and a data set of unresolved genotypes and then assigns to each individual sample one or more pairs of haplotypes that can resolve its genotype data; Assign lists every pair of haplotypes that can resolve an individual's genotype data. We input each ethnic group's data set individually with the 22 haplotypes. In this way we were able to determine which of the haplotypes suggested by EMHAPFRE were necessary for resolution of our genotype data, thus further refining the results. The genotype of every sample in this study could be accounted for by at least one pair of the 22 haplotypes predicted by EMHAPFRE from the complete data set. Six of the 22 haplotypes exist in all ethnic populations, and 11 of them are unique to a single population and hereafter are referred to as "private" haplotypes (table 2); each of these 11 haplotypes has a frequency of <1%.

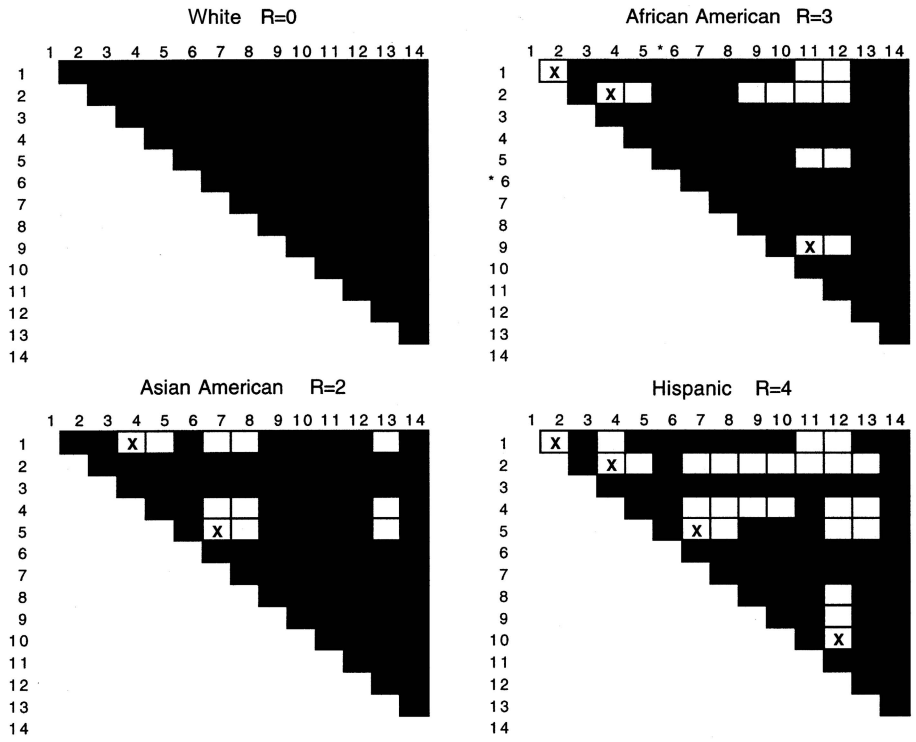

Figure 3 Four-gamete test for recombination in *ATM*. White boxes denote site pairs having four gametic types, which implies that recombination has occurred between these two sites. Also shown is the Hudson and Kaplan recombination statistic R, which is an estimate of the number and sites of recombination events needed to explain the results of the four-gamete matrix. A white box containing an " × " denotes a potential site of recombination. The asterisk (*) denotes an SNP that is not polymorphic in the sample population.

We analyzed primate DNA in order to approximate an ancestral *ATM* haplotype. Three haplotypes were found in 12 chromosomes (table 2). Two common chimpanzees, one bonobo, and three gorillas were genotyped by ASO hybridization and fluorescent sequencing; in cases in which ASO hybridization gave ambiguous results, fluorescent sequencing was used to confirm the genotype. None of the ape haplotypes was found among the 22 human haplotypes. One ape haplotype differs from a human haplotype by a single-base variant. This human haplotype is one of the least common (frequency .007) and occurs only in our African American study group. Only one of the human SNPs showed variation in the apes; the remainder were monomorphic. One common chimpanzee was heterozygous for IVS62+424g→a. The gorillas shared all but one allele with the chimpanzees. At IVS8−356t→c, gorillas are homozygous for a third allele (A), which is not found in either humans or chimpanzees.

Intragenic Recombination and LD

The small number of haplotypes seen in our study population suggests the possibility that recombination is reduced at the *ATM* locus. This is further evidenced by the results of the four-gamete test (fig. 3) (Hudson and Kaplan 1985). For a given haplotype AB, mutation may result in either Ab or aB. Haplotype ab arises only in the case of either recombination or repeat mutation. For the purpose of this analysis, we will consider repeat mutation to be rare and will use the four-gamete test as a measure of recombination. The four-gamete test was executed on unphased genotype data, in a pairwise fashion across SNP loci. This was done for each ethnic group separately. Interestingly, the four-gamete test found no site pairs with four gametes in the samples from white European Americans, implying a complete lack of recombination in that population. Low recombination was indicated for the other groups, as shown in figure 3.

Another test for recombination is that of Hudson and Kaplan (1985). Based on the resulting matrix of the four-gamete test, the Hudson and Kaplan parameter *R* is an estimate of the minimum number of recombination events in the history of the sample. For the white European American population, this estimate is 0 (fig. 3). For the other ethnic groups, *R* ranges from 4, in His-

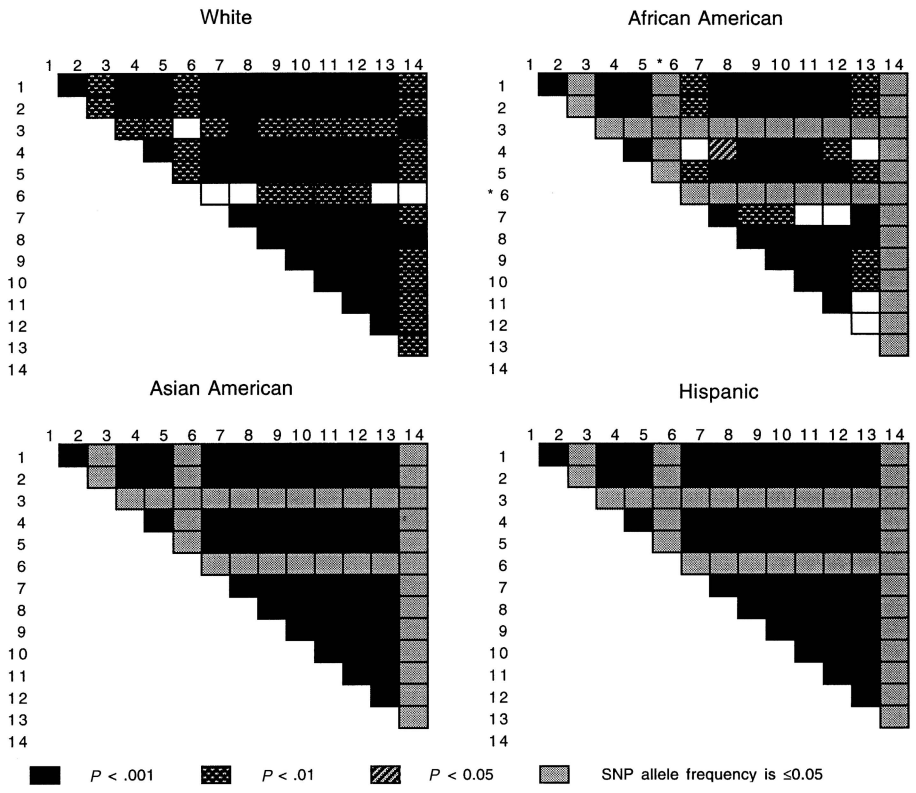

Figure 4 Fisher's exact test for LD in *ATM*. White boxes denote site pairs that do not have a significant value by Fisher's exact test, indicating linkage equilibrium. Gray columns and rows denote SNPs that have a minor-allele frequency ⩽.05. The asterisk (*) denotes an SNP that is not polymorphic in the sample population.

panics, to 2, in Asian Americans. The predicted sites of recombination are similar among ethnic groups. African Americans and Hispanics share two possible recombination sites in the 5′ end of the gene, and a third, in the 3′ end, could also be in the same location. The Asian American population shares one of the 5′ end sites and has, in the middle of the gene, another potential site of recombination, which is also present in Hispanics.

Further support for the hypothesis that there is minimal recombination at the *ATM* locus is provided by the results of Fisher's exact test (Weir 1996). We computed all possible pairwise comparisons between sites, to determine the degree of nonrandom association between sites. The majority of site pairs across all data sets show significance ($P < .001$), indicating that there is extensive disequilibrium at this locus (fig. 4). It has been demonstrated that alleles with frequencies ⩽.05 do not have the power for detection of disequilibrium (Lewontin 1995; Goddard et al. 2000). In this analysis, we included only SNPs having an allele frequency >.05. The Hispanic and Asian American populations were in complete disequilibrium. In the white European American population, the pattern of equilibrium followed the SNP with the lowest-frequency (.06) allele.

Disequilibrium was next measured by use of the statistic D', in a pairwise fashion across the 14 SNP loci (fig. 5). $D' = D/|D|\text{max}$, where $D = \text{p11} - \text{p1p2}$ and $|D|\text{max} = \text{max}(\text{p1p2},\text{q1q2})$ if $D < 0$ and $|D|\text{max} = \text{min}(\text{q1p2},\text{p1q2})$ if $D > 0$. D' ranges from 1 to −1, and this range is not influenced by allele frequency. A score of either 1 or −1 is considered to represent perfect disequilibrium. Interestingly, the results of this test are virtually superimposable on the results of the four-gamete test. The majority of site pairs are in perfect disequilibrium. The white European American population is in perfect disequilibrium across all sites. For the other groups, the sites with $|D'| < 1$ are exactly the same sites that have four gametes. We conclude that the *ATM* locus exhibits reduced recombination and extensive disequilibrium in all four ethnic groups, with the white European American population being the most extreme case.

Association Study

Ultimately, we aim to use these *ATM* haplotypes for association studies in populations with cancer. To evaluate the potential that these haplotypes have for identification of a particular mutation or polymorphism, we

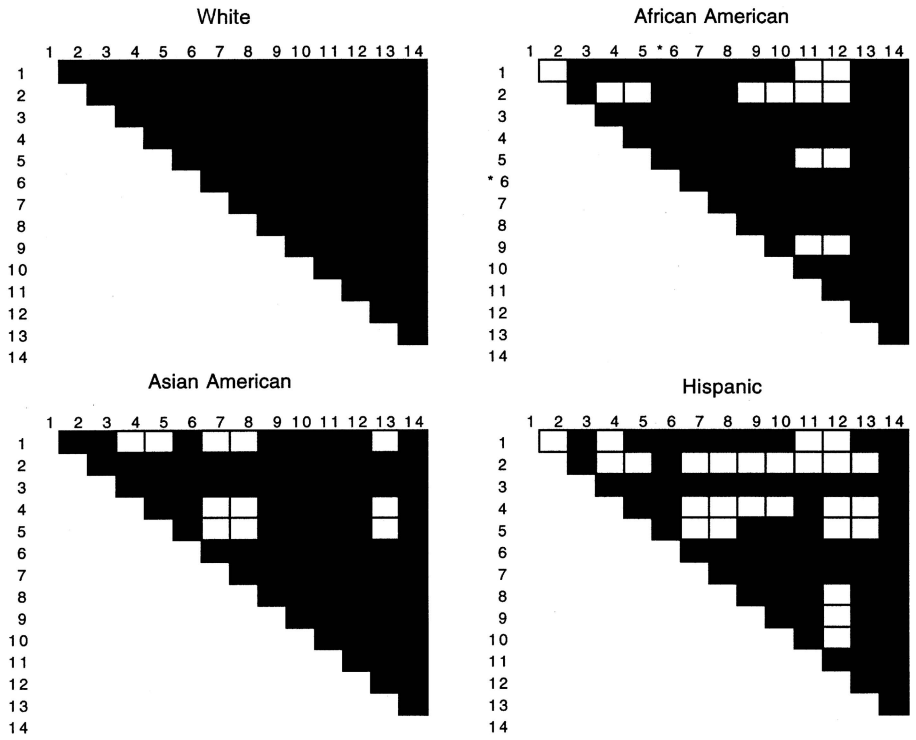

Figure 5 *D′*, measured as $D' = D/|D|\max$ in a pairwise fashion across 14 SNP loci. A score of either 1 or −1 is considered perfect disequilibrium. Black boxes denote site pairs with perfect disequilibrium; white boxes denote site pairs with $|D'| < 1$. The asterisk (*) pair denotes an SNP that is not polymorphic in the sample population.

performed a model association study. We tested the ability of these haplotypes to detect, by association, three different cSNPs in the *ATM* gene. These cSNPs were found by sequencing the reverse transcriptase–PCR products from *ATM* mRNA isolated from peripheral blood lymphocytes from cancer patients. cSNP1 is located in exon 4 and results in Ser49Cys. Positioned in exon 38, cSNP2 results in Asp1853Asn; and cSNP3, which results in Pro1054Arg, is located in exon 23. A population of 941 individuals was screened for these three cSNPs, by ASO hybridization. The resulting frequencies of the cSNPs in this population are shown in table 3.

In the model association study, samples from white European Americans in the 941-individual collection that were found to possess one of the three cSNPs were considered to be the "case" population. The "control" population consisted of samples from white European Americans from the same collection that were randomly chosen and negative for the cSNPs; because of the low frequency of these cSNPs in other ethnic groups, only samples from white European Americans were used in this association study. All case and control samples were genotyped for the 14 *ATM* neutral sequence variants, via ASO hybridization. To assign haplotypes to individual samples, we used Assign and the initial 22 *ATM* haplotypes.

Each cSNP showed a significant association with a different, specific *ATM* haplotype (table 4). cSNP1 showed an association with haplotype 2, cSNP2 with haplotype 15, and cSNP3 with haplotype 17. Haplotype 2 was present at a frequency of .29 in the control population (no. of chromosomes [*c*] = 152) and at a frequency of .64 in the cSNP1 population ($c = 14$); haplotype 15 was present at a frequency of .07 in the control population ($c = 112$) and at a frequency of .57 in the cSNP2 population ($c = 56$); and haplotype 17 was pre-

Table 3

Frequencies for Three *ATM* cSNPs in the Control Population

	Frequency in 941 Individuals			
	White European Americans	African Americans	Asian Americans	Hispanic Americans
cSNP1	.005	.000	.001	.001
cSNP2	.066	.001	.002	.017
cSNP3	.015	.001	.000	.005

Table 4

Association of *ATM* Haplotypes and *ATM* cSNPs in Individual "Case" and Control Populations

		Frequency in[a]		*P* for	
	Haplotype	Control Population	"Case" Population	Genotype Association[b]	Haplotype Association[c]
cSNP1	2	.29 (c = 152)	.64 (c = 14)	.0478	.0166
cSNP2	15	.07 (c = 112)	.57 (c = 56)	.0000	.0000
cSNP3	17	.08 (c = 146)	.52 (c = 54)	.0000	.0000

Note.—Samples carrying one of three *ATM* cSNPs were genotyped, by ASO hybridization, for the 14 *ATM* noncoding SNPs.

[a] Each cSNP was found to occur on separate *ATM* haplotypes. c = total number of white European American chromosomes genotyped.

[b] By 2 × 2 contingency table.

[c] By 3 × 3 contingency table.

sent at a frequency of .08 in the control population (c = 146) and at a frequency of .52 in the cSNP3 population (c = 54). These are 2-fold, 8-fold, and 6.5-fold increases in the frequencies of haplotypes 2, 15, and 17, respectively; and the *P* values for these associations are .0166, .0000, and .0000, respectively. Genotype correlations were also present, with *P* values of .0478, .0000, and .0000, respectively (table 4).

One of the great challenges in studying the genetics of a complex disease such as cancer is its multifactorial etiology. As presented in table 4, the data from our simulated association study model a scenario in which all "cases" are caused by a single mutation. To more accurately simulate an association study with a complex disease, we reanalyzed our data. We considered the three groups of samples carrying the variant cSNPs as one "case" population. In this analysis, two of the three haplotypes that originally had shown an association demonstrated a significant increase in frequency (table 5). No increase in frequency was apparent for haplotype 2, which had previously shown a twofold increase in the cSNP1 population; haplotype 15 showed a fourfold increase (P = .0002); and haplotype 17 showed a threefold increase (P = .0002). Thus, we successfully demonstrated the ability of these *ATM* haplotypes to discern members of our case population who carry a particular SNP. The results of these studies indicate a significant potential for the use of haplotypes extending over a large genomic region, to detect disease associations through a case-control–study design in a general population.

Discussion

In this study, we have presented a strategy for uncovering the genetic contribution to complex disease. Specifically, we have demonstrated the utility of a complex SNP-based–haplotype approach to association studies and have detected significant LD at the *ATM* locus, extending ~142 kb. The results of this study provide proof of principle for the use of SNP-haplotype data in the detection of genetic factors contributing to complex disease.

We sequenced 13.5 kb of the *ATM* gene in five unrelated individuals and detected 17 SNPs in noncoding regions. We then utilized these neutral sequence variants spanning 142 kb of the *ATM* gene to construct haplotypes for this genomic locus. The expectation-maximization algorithm EMHAPFRE (Excoffier and Slatkin 1995) was used to predict haplotypes from genotype data on 295 individuals from four ethnic groups. Twenty-two haplotypes and their frequencies were predicted by EMHAPFRE, for the total population. Three of these 22 haplotypes have a frequency of ⩾10%. This concurs with the findings of Li et al. (1999), who also used neutral sequence variants to detect three major haplotypes at the *ATM* locus. Six of the 22 haplotypes exist in all four ethnic populations in our study and are also the most commonly occurring haplotypes. There are 11 private haplotypes, each of which has a frequency of <1%.

We verified the reliability of the haplotype-prediction algorithm by using several tests. First, we genotyped individuals from nine three-generation CEPH families (n = 87). This allowed us to determine haplotypes by inspection of allele segregation. The CEPH genotype data were also analyzed by EMHAPFRE, and the resulting haplotypes agreed completely with those in-

Table 5

Association of *ATM* Haplotypes and *ATM* cSNPs in Combined Case Population

	Frequency in Combined Case Population	*P* for	
Haplotype	(c = 124)	Genotype Association	Haplotype Association
2	.19	.9644	.7606
15	.27	.0000	.0002
17	.24	.0000	.0002

Note.—See footnotes to table 4.

ferred on the basis of transmission data. Next, we used Assign, a script written in Microsoft Excel Visual Basic, to assign pairs of haplotypes to individual genotypes. Given the 22 haplotypes predicted by EMHAPFRE, Assign successfully resolved the genotype data for all 295 individuals in this study. The results of EMHAPFRE were tested against another haplotype-prediction program, one that does not use the expectation-maximization algorithm and that does not assume that Hardy-Weinberg is in effect. This program, termed "Data Mining," uses the resulting matrix of the four-gamete test to inform the process of haplotype prediction so that recombination may influence outcome (N. Wang, R. Chakraborty, M. Kimmel, and L. Jin, personal communication). There were minor differences in the results of this comparison. For the population of white European Americans, the outcome of each program was identical. This is not surprising, since the four-gamete test reveals no evidence for recombination in this population. The results of these trials confirm that EMHAPFRE was successful in estimating the correct haplotypes necessary to sufficiently resolve our data set. We feel confident that the size and diversity of our data set has allowed us to describe in relative depth the haplotype architecture of *ATM*. Consequently, we have chosen to use, as the foundation for further studies, the 22 haplotypes predicted from the complete data set.

With a minimal amount of sequencing (13.5 kb in five individuals), we were able to detect highly informative neutral sequence variants spanning a large genomic region. In sequencing 10 chromosomes form white European Americans, we found SNPs that have a common occurrence in four different ethnic groups. In all ethnic groups, the majority (11 of 14) of SNPs identified in this study have a minor-allele frequency of $\geqslant$25%. SNPs with frequencies in the range of .2–.5 have the highest information content for association and LD studies (Kruglyak 1997). Although most SNPs had a high minor-allele frequency in all ethnic groups, allele frequencies varied across ethnic groups. This is in accordance with several other studies that have found population differences in SNP-allele frequencies (Lai et al. 1998; Nickerson et al. 1998; Cargill et al. 1999; Halushka et al. 1999; Goddard et al. 2000). Variations in allele frequencies are most pronounced in the African American population. Four SNPs (IVS21−77t→c, IVS55+186c→t, IVS62+424g→a, and IVS62−973a→c) have a minor-allele frequency that is reduced by 40%–75% in African Americans, compared with that in other ethnic groups. A fifth SNP, IVS46−257a→c, was not found in the African American samples. These differences illustrate that there is population structure in SNP-allele frequencies that is an important factor to consider when SNP-based association and LD studies are designed.

Comparison of genotype data from six great apes was instructive for approximating ancestral haplotypes and SNP alleles. Genotyping revealed three haplotypes in this population, none of which is identical to the human *ATM* haplotypes. Of the 14 SNPs, 2 showed variation in the ape population. One common chimpanzee was heterozygous for IVS62+424g→a, and all three gorillas were homozygous for a third allele (A) at IVS8−356t→c. The extent of homozygosity in this sample indicates that most of the SNPs found varying in the human population have arisen since the divergence of the human lineage from the last common ancestor shared with the chimpanzee. This agrees with the assertion by Hacia et al. (1999)—that is, that most current neutral human polymorphisms are not shared with the chimpanzee (Hacia et al. 1999). It may also imply that these SNPs are not hypermutable sites, since more variation might be expected in the 12 primate chromosomes analyzed. Although these SNPs are common in man, they are not due to hypermutability; rather, they are old enough to be found throughout diverse ethnic groups.

The results of this study show a remarkable lack of recombination at the *ATM* locus. This effect is most profound in the white European American population, in which no evidence for recombination is detected by the four-gamete test and in which D' shows perfect disequilibrium across all SNPs. Low recombination is implicated for the African American, Asian American, and Hispanic groups as well. The possibility of low recombination was suspected on the basis of the seemingly small number of haplotypes found at this locus. Twenty-two haplotypes with 14 loci is not considerably greater than the $n + 1$ (i.e., 15) that would be expected if there is no recombination. Another study, performed in parallel with this one, used the same approach as that described here and serves as a direct comparison: D. Trikka, Z. Fang, A. Renwick, S. Jones, R. Chakraborty, M. Kimmel, and D. L. Nelson (unpublished data) used neutral sequence variants dispersed across the BLM, WRN, and RECQL loci, to derive haplotypes for these regions; their study used the same sample population, with fewer SNPs (8, 13, and 11 respectively) for haplotype construction, and found considerably larger numbers of haplotypes (50, 56, and 47, respectively) at each locus. The key difference between these loci and *ATM* is the amount of recombination and LD reported. Trikka et al. found more evidence for recombination and linkage equilibrium when the four-gamete test and Fisher's exact test were used. For *ATM,* the four-gamete test revealed few site pairs with four gametes. The Hudson-Kaplan recombination statistic R ranged from 0, in white European Americans, to 4, in Hispanics. Analysis by both Fisher's exact test and D' indicated extensive LD for *ATM,* in all ethnic groups studied. Figure 5

shows extensive disequilibrium, with >72% of site pairs having perfect disequilibrium in all ethnic groups.

Using a model association study, we have successfully demonstrated the ability of *ATM* haplotypes to identify chromosomes carrying specific coding polymorphisms. The three cSNPs that we used as candidates for detection had varying frequencies in our control population of white European Americans (cSNP1, .005; cSNP2, .066; and cSNP3, .015). When each of the three cSNP populations was analyzed individually, each cSNP showed a significant association with a different *ATM* haplotype. cSNP1 showed an association with haplotype 2, cSNP2 with haplotype 15, and cSNP3 with haplotype 17 ($P = .0166$, .0000, and .0000, respectively); the increase in haplotype frequency in cases versus controls was 2-fold, 8-fold, and 6.5-fold, respectively. To model the potential for multiallelic etiology of a complex disease, we combined the three populations of samples carrying the cSNPs into one "case" population. In this analysis, two haplotypes demonstrated a readily detectable increase in frequency: haplotype 15 showed a fourfold increase, and haplotype 17 showed a threefold increase in frequency; no frequency increase was apparent for haplotype 2, which had previously shown a twofold increase in the cSNP1 population.

The association that becomes undetectable (i.e., haplotype 2 with cSNP1) involves the haplotype occurring most commonly (frequency .29) in the general population. Haplotype 15 shows the greatest increase in frequency and is the least common of the three haplotypes, with a control frequency of .05. This leads us to an important point for future association studies. Haplotypes with lower frequencies in control populations may be more effective for detection of associations. However, it is important to note that haplotype 17, which is the third most frequent haplotype (frequency .10), nevertheless showed a 2.6-fold increase in frequency in the combined cSNP population. An additional factor contributing to detection in this study is frequency of the mutation. In the case of cSNP1 and haplotype 2, in which the association becomes undetectable, the most frequent haplotype was associated with the least common SNP (cSNP1, .006). The difference in frequency between cSNP1 and the other cSNPs is a factor of 10. Both the haplotype frequency and the cSNP frequency contribute to detection. This underscores the idea that several factors, including frequency of haplotype, frequency of mutation, and age of mutation, contribute to limits of detectability.

This model association study demonstrates proof of principle for the use of complex SNP haplotypes covering candidate genes, in the detection of genetic factors contributing to complex disease. We have successfully demonstrated the ability of these *ATM* haplotypes to discern members of our "case" population that carry a particular coding SNP. The results of these studies indicate that haplotypes extending over a large genomic region have a significant potential for detection of disease associations.

There is much interest in the use of SNPs in genomewide association studies and other LD-based strategies. Our approach and analyses bear on those strategies, in several regards. First, LD estimates from simulation studies have been as low as 3 kb of meaningful LD (Kruglyak 1999). This calculation suggests that a very-high-density map with as many as 0.5–3 million SNPs would be necessary for effective association studies (Kruglyak 1999). Our results and those of other studies (Collins et al. 1999; Eaves et al. 2000; Moffatt et al. 2000; Taillon-Miller et al. 2000) indicate, to the contrary, that significant LD can be found extending as far as several hundred kilobases. This should reduce the number of SNPs necessary for genomewide linkage studies. Comparison of LD at *ATM* versus the results of the *LPL* study (Clark et al. 1998) in which LD patterns were complex over just 9.7 kb supports the idea that LD varies widely throughout the genome, indicating that some regions will require SNPs that are more densely spaced. Second, higher-frequency (.2–.5) SNPs are more robust, whereas rare SNPs may be less useful and, in some analyses, may confound results. More than half of the SNPs used to construct haplotypes in the *LPL* study had a relative allele frequency of <.2. This resulted in 67 of 71 individuals having a unique haplotype. By using fewer markers (14) with higher frequency (.20), we were able to effectively elucidate the haplotype architecture and the LD and recombination profiles for the *ATM* genomic locus (142 kb). These haplotypes were used successfully in association studies, to detect coding polymorphisms in the *ATM* gene. We conclude that reasonably spaced, highly informative SNPs have the ability to define a larger number of ancestral chromosomes and have increased power for population-based association studies.

Acknowledgments

The authors thank Ranajit Chakraborty of the University of Texas Health Science Center (Houston) and Marek Kimmel of Rice University (Houston), for discussion and review of the manuscript; Jason Deats, for his valued contribution to the design and code for Assign; and Melissa Bondy, Alice Sigurdson, and Scott Manatt, of M.D. Anderson Cancer Center, for collection of normal DNAs. P.E.B was supported by a fellowship from the W.M. Keck Center for Computational Biology (Houston) (funded by U.S. National Library of Medicine Training Grant 1T15LM07093). T.A.B. was supported by Department of Defense Breast Cancer Research Program Career Development Award CBC980154. This work was supported by the Kleburg Fund for New and Innovative Research and by U.S. National Cancer Institute grant CA75432.

Appendix A

Primers Used for PCR and Sequencing

f1.atm, ATGGTCATCTCGTTACAGGCAATGC
r1.atm, CCCCAAGTGACTGAAGGCATCTAGG
f2.atm, TGGTGGAACCATTTCCGTTTAACG
r2.atm, GCGCCCTTCTAATAACCCGCC
f3.atm, GCCCAGAACCTCCGAATGACG
r3.atm, CGACTTAGCGTTTGCGGCTCG
f4.atm, TGGCTGGCAACATTACCAACTGC
r4.atm, TGCATCTTTTTCTGCCTGGAGGC
f5.atm, TGTGTGCTAGGGAGGAATCTGGTGG
r5.atm, GGCTGTCTCTAGGCTTGTTGAGGGC
f6.atm, CCATCATCCGAAAGGAGCCAAAAC
r6.atm, GCAGCAATTTCCCTGTTTCTGCC
f7.atm, AAATTGGCAGGATGATGAGGATGC
r7.atm, GCTGTCAAGCTGCATCAGCGTTAG
f8.atm, CCAAAGCGTGCCAGAATGGTATG
r8.atm, CCAAAGCGTGCCAGAATGGTATG
f9.atm, GGTATGCGTAGCGGGGCTAGTGAG
r9.atm, CGCAGGAAAAAGCCAGATGCAATC
f10.atm, GCCCTAGCCCCAGTGTATGTGGAG
r10.atm, GGCAGCCAGTTTCCGAGAACTACC
f11.atm, TTTTTGGCAAGGTGAGTATGTTGGC
r11.atm, TGCGAACTTGGTGATGATTGTCAGC
f12.atm, AGATTGTTCCAGGACACGAAGGGAG
r12.atm, TTTCTTCCCATTGTCACCTGTTCCC
f13.atm, TGCGAAAAACAGGCTTTGTTTGC
r13.atm, GGTGATGGAAAAGAGACGGGGC
f14.atm, GCAAGTCCCCTCACCAGCAACAC
r14.atm, GATGCCTTCCCATCATCCTGATACC
f15.atm, TCTGGGAAGAAGTTACGCAGGGAAC
r15.atm, CTGACTGGCACTAGAATTTGCTGGC
f16.atm, GGCGGAATGAATGTGAGTTATGCG
r16.atm, CCAGGTGATTTCTCCATCCCGTG
f17.atm, CTGCCTAAAGCAGCAGTTTTTGCC
r17.atm, TGTTGCTATCCCGAAGCTGAAACC
f18.atm, GGTGTGTAAGCAAGAATGCCTGGG
r18.atm, GCCACAGATTTTGAGACCACTGCAC
f19.atm, TAGTTTGTATGGCTGTGGTGGAGGG
r19.atm, CATCCCTCTGCTTCAGGAGTATCCC
f20.atm, CCAGTAGGGGGTCCCTCATTTCC
r20.atm, TGAGAAGCTGGGAGTGTTTCTGCC
f21.atm, CCCCGTACATGAAGGGCAGTTG
r21.atm, TGGGTGGCTGGGCTAATGAAGAG
f22.atm, GGTTCAGCGAGAGCTGGAGTTGG
r22.atm, GCAGCAGGGGGAAAACCCAC
f23.atm, CCACAGATTAGCAACAAGTTGGGGC
r23.atm, TGGCATAAGCACACGGAAACTCTCC
f24.atm, AGGTTCCGATGGCAAGGAGAGG
r24.atm, CTGTGTCTTTCCACCACTCCCCAG
f25.atm, CAGTCATGGTTCTGGGGAGAGAAGC
r25.atm, GCCTTTCTGATTTCCCTTCCTGCC
f26.atm, CTTGATGGTGGGAGGGACTTAGGG
r26.atm, TGCCTAGATGTTTGAGAGCCTGCC
f27.atm, CAGGGCACACAGGGTACAGTGTAGG
r27.atm, TCAGTTCAGACCATCTCATGCCTCC
f28.atm, CAGGGGGATGATAGTGATGATGTGG
r28.atm, TTCAAAACATACATGCCCTGCCTTC
f29.atm, CAAAGACTGAGAGCTGAGCCCAGTG
r29.atm, GCACAATCTCCTCCTTTCTGCTGC
f30.atm, TGGTTTAGAAATGCCTTCAGCCCC
r30.atm, TGCACTCTACCTGCCATGCTTCC
f31.atm, GCCATGTCAGTGCCCAACTTGAAG
r31.atm, TTGGTGCTGCGTTTGGAATCTTG
f32.atm, GATTCCAAACGCAGCACCAAACC
r32.atm, GGTAGTTGATGGGGGAGGGGAAC
f33.atm, GTTCCCCTCCCCCATCAACTACC
r33.atm, GAGCACAGTGCCTTCTTCCACTCC
f34.atm, CCCTGACAATCTGGGGCACAAAC
r34.atm, CCGTGGCTTTTGCTGGCATTC
f35.atm, GTCCTGTGGCATTGTGCATAACTCC
r35.atm, GCAGACATTAGGCATAAGCCCCTTC
f36.atm, GATGACTGCCCTTGTTCCCCAAG
r36.atm, TGGTTAAGTTGCTTTTCCCCCCAG

Appendix B

Primer Sequences and Concentrations Used for Multiplex PCR

Group 8:
3F *ATM*, 5′-GCCCAGAACCTCCGAATGACG-3′; and 3R-2 *ATM*, 5′-GCCGTGAAGCGAAAGAGGCG-3′ (0.25 μM)
11F *ATM*,5′-TTTTTGGCAAGGTGAGTATGTTGGC-3′; and 11R *ATM*, 5′-TGCGAACTTGGTGATGATTGTCAGC-3′ (0.25 μM)
14F *ATM*, 5′-GCAAGTCCCCTCACCAGCAACAC-3′; and 14R *ATM*, 5′-GATGCCTTCCCATCATCCTGATACC-3′ (0.25 μM)
23F-2 *ATM*, 5′-GGTGGAATCTGGTCTAGTTACCC-3′; and 23R *ATM*, 5′-TGGCATAAGCACACGGAAACTCTCC-3′ (0.25 μM)
27F *ATM*, 5′-CAGGGCACACAGGGTACAGTGTAGG-3′; and 27R *ATM*, 5′-TCAGTTCAGACCATCTCATGCCTCC-3′ (0.188 μM)
29R *ATM*, 5′-CAAAGACTGAGAGCTGAGCCCAGTG-3′; and 29R *ATM*, 5′-GCACAATCTCCTCCTTTCTGCTGC-3′ (0.125 μM)
30F *ATM*, 5′-TGGTTTAGAAATGCCTTCAGCCCC-3′; and 30R-2 *ATM*, 5′-CAGCCAGTCCAACATAAATCAG-3′ (0.25 μM)
31F *ATM*, 5′-GCCATGTCAGTGCCCAACTTGAAG-3′; and 31R *ATM*, 5′-TTGGTGCTGCGTTTGGAATCTTG-3′ (0.25 μM)
Group 7:
7R *ATM*, 5′-GCTGTCAAGCTGCATCAGCGTTAG-3′; and 7F-2 *ATM*, 5′-GTTGGATTACCATGTTCACCAG-3′ (.188 μM)
10F *ATM*, 5′-GCCCTAGCCCCAGTGTATGTGGAG-3′and 10R-2 *ATM*, 5′-GCAGAGATAATCATGGGCAGG-3′ (0.25 μM)
15F *ATM*, 5′-TCTGGGAAGAAGTTACGCAGGGAAC-3′; and 15R-2 *ATM*, 5′-TGGGGAGACTATGGTAAAAGAGG-3′ (0.31 μM)
16F *ATM*, 5′-GGCGGAATGAATGTGAGTTATGCG-3′; and 16R *ATM*, 5′-CCAGGTGATTTCTCCATCCCGTG-3′ (0.25 μM)
20F *ATM*, 5′-CCAGTAGGGGGTCCCTCATTTCC-3′; and 20R *ATM*, 5′-TGAGAAGCTGGGAGTGTTTCTGCC-3′ (0.25 μM)

25F *ATM,* 5′-CAGTCATGGTTCTGGGGAGAGAAGC-3′; and 25R-2 *ATM,* 5′-CTATCAATATCTAGCTCTGGGGC-3′ (0.15 μM)
28F *ATM,* 5′-CAGGGGGATGATAGTGATGATGTGG-3′; and 28R *ATM,* 5′-TTCAAAACATACATGCCCTGCCTTC-3′ (0.5 μM)

Appendix C

Probes Used for ASO Hybridization

*ATM*Aso 3T, 5′-TAACCCTCCTTCCCGC-3′
*ATM*Aso 3a, 5′-TAACCCTCCATCCCGC-3′
*ATM*Aso 7T, 5′-AAGGAACTTGTAATATTTTTC-3′
*ATM*Aso 7c, 5′-AGGAACTCGTAATATTTTTC-3′
*ATM*Aso 10T, 5′-TGGGAAACATGACCAGGG-3′
*ATM*Aso 10c, 5′-GGGAAACACGACCAGGG-3′
*ATM*Aso 11T, 5′-GTAACTTATAATAACCTTTC-3′
*ATM*Aso 11c, 5′-GAAGTAACTTACAATAACC-3′
*ATM*Aso 14C, 5′-TCTGTACAAGAAAAATTTG -3′
*ATM*Aso 14g, 5′-TCTGTAGAAGAAAAATTTG-3′
*ATM*Aso 15C, 5′-TTTCCTCTCAGTCTACAGG-3′
*ATM*Aso 15t, 5′-TTTTTCCTCTTAGTCTACAGG-3′
*ATM*Aso 16C, 5′-TAGAGATGATGTCGGCTTC-3′
*ATM*Aso 16t, 5′-CTAGAGATGATGTTGGCTTC-3′
*ATM*Aso 20A, 5′-GTAATGTCAGAGTATTAAA-3′
*ATM*Aso 20c, 5′-TAATGTCAGCGTATTAAA-3′
*ATM*Aso 23T, 5′-CAAAAGCTTCTCTTGCTTT-3′
*ATM*Aso 23c, 5′-AAAAGCTTCTCCTGCTTTC-3′
*ATM*Aso 25C, 5′-TTTTTTGTGGCATTCACAC-3′
*ATM*Aso 25t, 5′-TTTTTTGTGGTATTCACAC-3′
*ATM*Aso 27C, 5′-CTGCTCATGCCTCCTCTC-3′
*ATM*Aso 27t, 5′-CTGCTCATGTCTCCTCTCC-3′
*ATM*Aso 28C, 5′-TTCTATTAAACAGTATTA-3′
*ATM*Aso 28a, 5′-TTCTATTAAAAAGTATTA-3′
*ATM*Aso 29.1T, 5′-GATAAAGATATGTTGACAA-3′
*ATM*Aso 29.1c, 5′-GATAAAGATACGTTGACAA-3′
*ATM*Aso 29.2C, 5′-ACTTCCTGACGAGATACAC-3′
*ATM*Aso 29.2t, 5′-ACTTCCTGATGAGATACAC-3′
*ATM*Aso 30c, 5′-CCTAAGCCACGTTCCTCTA-3′
*ATM*Aso 30t, 5′-CCTAAGCCATGTTCCTCTA-3′
*ATM*Aso 31.1C, 5′-AAATAGAGCGATTTTGGTT-3′
*ATM*Aso 31.1t, 5′-AAATAGAGAGATTTTGGTTC-3′
*ATM*Aso 31.2C, 5′-AGAAATTCCTCATGAACTC-3′
*ATM*Aso 31.2a, 5′-AGAAATTCATCATGAACTC-3′

Electronic-Database Information

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

GenBank Overview, http://www.ncbi.nlm.nih.gov/Genbank Overview.html (for genomic sequence [accession number U82828])

References

Boerwinkle E, Ellsworth DL, Hallman DM, Biddinger A (1996) Genetic analysis of atherosclerosis: a research paradigm for the common chronic diseases. Hum Mol Genet 5:1405–1410

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238 (erratum: Nat Genet 23:373 [1999])

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173–15177

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990) Centre d'Étude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

DeMarchi JM, Richards CS, Fenwick RG, Pace R, Beaudet AL (1994) A robotics-assisted procedure for large scale cystic fibrosis mutation analysis. Hum Mutat 4:281–290

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet 25:320–323

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Gatti RA, Boder E, Vinters HV, Sparkes RS, Norman A, Lange K (1991) Ataxia-telangiectasia: an interdisciplinary approach to pathogenesis. Medicine (Baltimore) 70:99–117

Gatti RA, Tward A, Concannon P (1999) Cancer risk in ATM heterozygotes: a model of phenotypic and mechanistic differences between missense and truncating mutations. Mol Genet Metab 68:419–423

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, Collins FS (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 22:164–167

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of

single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239–247

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Josefsson AM, Magnusson PK, Ylitalo N, Quarforth-Tubbin P, Ponten J, Adami HO, Gyllensten UB (1998) p53 polymorphism and risk of cervical cancer. Nature 396:531–532

Kraft HG, Windegger M, Menzel HJ, Utermann G (1998) Significant impact of the +93 C/T polymorphism in the apolipoprotein(a) gene on Lp(a) concentrations in Africans but not in Caucasians: confounding effect of linkage disequilibrium. Hum Mol Genet 7:257–264

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24

——— (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22: 139–144

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

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539

Lewontin RC (1995) The detection of linkage disequilibrium in molecular sequence data. Genetics 140:377–388

Li A, Huang Y, Swift M (1999) Neutral sequence variants and haplotypes at the 150 Kb ataxia-telangiectasia locus. Am J Med Genet 86:140–144

MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, Taylor S, Allitto B, Altherr M, Myers R, Lehrach H, Collins F, Wasmuth J, Frontali M, Gusella J (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet 1:99–103

Moffatt MF, Traherne JA, Abecasis GR, Cookson WO (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. Hum Mol Genet 9:1011–1019

Morrell D, Chase CL, Swift M (1990) Cancers in 44 families with ataxia-telangiectasia. Cancer Genet Cytogenet 50:119–123

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

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

Rotman G, Shiloh Y (1999) ATM: a mediator of multiple responses to genotoxic stress. Oncogene 18:6135–6144

Savitsky K, Bar-Shira A, Gilad S, Rotman G, Ziv Y, Vanagaite L, Tagle DA, et al (1995) A single ataxia telangiectasia gene with a product similar to PI-3 kinase. Science 268:1749–1753

Stankovic T, Kidd AM, Sutcliffe A, McGuire GM, Robinson P, Weber P, Bedenham T, Bradwell AR, Easton DF, Lennox GG, Haites N, Byrd PJ, Taylor AM (1998) ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer. Am J Hum Genet 62: 334–345

Storey A, Thomas M, Kalita A, Harwood C, Gardiol D, Mantovani F, Breuer J, Leigh IM, Matlashewski G, Banks L (1998) Role of a p53 polymorphism in the development of human papillomavirus-associated cancer. Nature 393:229–234

Strittmatter WJ, Roses AD (1995) Apolipoprotein E and Alzheimer disease. Proc Natl Acad Sci USA 92:4725–4727

Swift M, Morrell D, Cromartie E, Chamberlin AR, Skolnick MH, Bishop DT (1986) The incidence and gene frequency of ataxia-telangiectasia in the United States. Am J Hum Genet 39:573–583

Swift M, Morrell D, Massey RB, Chase CL (1991) Incidence of cancer in 161 families affected by ataxia-telangiectasia. N Engl J Med 325:1831–1836

Swift M, Reitnauer PJ, Morrell D, Chase CL (1987) Breast and other cancers in families with ataxia-telangiectasia. N Engl J Med 316:1289–1294

Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet 25:324–328

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA

Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, Nakura J, Miki T, Ouais S, Martin GM, Mulligan J, Schellenberg GD (1996) Positional cloning of the Werner's syndrome gene. Science 272:258–262