Complex SNP-Based Haplotypes in Three Human Helicases: Implications for Cancer Association Studies

Dimitra Trikka,[1] Zhe Fang,[1] Alex Renwick,[3] Sally H. Jones,[1] Ranajit Chakraborty,[2] Marek Kimmel,[3] and David L. Nelson[1,4]

[1]*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA;* [2]*Center for Genomic Information, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio 45221, USA;* [3]*Department of Statistics, Rice University, Houston, Texas 77005, USA*

We have initiated a candidate gene approach to study variation and predisposition to cancer in the four major ethnic groups that constitute the U.S. population (African Americans, Caucasians, Hispanics, and Asians). We resequenced portions of three helicase genes (*BLM*, *WRN*, and *RECQL*) identifying a total of 37 noncoding single nucleotide polymorphisms (SNPs). Haplotype inference predicted 50 haplotypes in *BLM*, 56 in *WRN*, and 47 in *RECQL* in a sample of 600 chromosomes. Approximately 10% of the predicted haplotypes were shared among all ethnic groups. Linkage disequilibrium and recombination effects showed that each locus has taken a diverse evolutionary path. Primate DNA analysis of the same loci revealed one human haplotype per gene shared with the great apes, indicating that the observed diversity occurred since the divergence of humans from the last common ancestor. In *BLM*, we confirmed the presence of a founder haplotype among Ashkenazi Jews homozygous for the blm^{Ash} mutation. The cosegregating haplotype was seen in all (6/6) samples of Ashkenazi descent, whereas in the general population it has a low frequency (0.02) and was not found in African Americans. In *WRN*, ethnic samples were studied for their haplotype content and the presence or absence of six previously described coding SNPs (cSNPs). Hispanic individuals carrying two of these cSNPs showed a 60% increase in the frequency of a common haplotype (haplotype No. 28). In the pooled sample, no association was found. Because (1) the majority of the haplotypes are population specific and (2) the patterns of linkage disequilibrium, recombination, and haplotype diversity are markedly different between gene regions, these data show the importance of either ethnically matched controls or within-family-based disease-gene association studies.

[The sequence data described in this paper have been submitted to the GenBank data library under accession no. AC006559. Online supplemental material available at http://www.genome.org]

With the completion of a human reference sequence through the Human Genome Project, the ability to define the role of genomic variation in disease risk is at hand. A starting point is to assess the type and distribution of sequence variation within and between ethnic populations. In recent years, single nucleotide polymorphisms (SNPs) have been favored as more tractable genotypic markers. SNPs, although reduced in allelic diversity, have a number of advantages over microsatellites, such as abundance (one every 750–1000 bp; Kwok et al. 1996; Wang et al. 1998), stability, and the capacity for highly automated analysis. As a consequence, SNPs are being utilized extensively as markers of choice in human genetic studies ranging from comparative population variation (Halushka et al. 1999; Goddard et al. 2000; Kidd et al. 2000; Ober et al. 2000) to disease linkage studies (Kruglyak 1997, 1999; Lai et al. 1998; Martin et al. 2000).

Several investigators (Spielman et al. 1993; Risch and Merikangas 1996; Long and Langley 1999) have suggested that association-based studies are more powerful than linkage studies in identifying genes or variants that contribute to complex trait phenotypes. This is based on the assumption that one of the markers used in the analysis will in fact be the polymorphism contributing to the disease variation or that neutral markers will be in linkage disequilibrium with the disease-causing site. These assumptions led to suggestions of genome-wide (Risch and Merikangas 1996) or candidate-gene scans (Halushka et al. 1999; Long and Langley 1999; Martin et al. 2000) using dense sets of SNPs as the means to identify sequences contributing to variation in complex traits.

In this study we set out to test the use of common variants and their associated haplotypes for detecting functional variants contributing to cancer risk. Given that numerous genetic lesions are found in tumors, along with the fact that carcinogenesis requires several somatic hits (Knudson 1996), we have begun a candidate-gene approach and concentrated our efforts on detecting variation in genes that play a role in DNA maintenance and chromosomal integrity. We have utilized primarily noncoding sequence variants. The rationale for this choice was to avoid bias that might be introduced by studying sites with functional effects. Our goal was to build haplotypes capable of marking chromosomes carrying common and functionally significant variants. However, we pre-

[4]Corresponding author.
E-MAIL nelson@bcm.tmc.edu; FAX (713) 798-5386.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.176702.

ferred not to assume the type of variant that might be relevant. Therefore, exclusive focus on relatively rare coding sequence changes might not allow common, functionally relevant intronic or promoter variants to be uncovered. A further benefit of such an approach is that it allows a more complete analysis of the population history of the loci under study (Bonnen et al. 2000).

BLM, *WRN*, and *RECQL* are genes encoding proteins with significant similarity with the *Escherichia coli* RecQ-type DNA helicases (Puranam and Blackshear 1994; Ellis et al. 1995; Yu et al. 1996) involved in duplex DNA unwinding (Umezu et al. 1990; Puranam and Blackshear 1994; Gray et al. 1997; Karow et al. 1997). Although to date no disease has been associated with *RECQL*, homozygous loss of function in *BLM* and *WRN* result in two distinct syndromes, Bloom's and Werner's, respectively (Ellis et al. 1995; Foucault et al. 1997; Moser et al. 1999). Both syndromes exhibit high tumor incidence attributed to high levels of chromosome instability and somatic mutation (Goto et al. 1981, 1996; Fukuchi et al. 1989; German 1993). Variation in these genes in the general population may play a role in predisposition to tumor formation or progression.

We carried out a pilot study of functionally neutral variation in the three helicases in four ethnic populations (Caucasians, African Americans, Hispanics, and Asians). We employed a resequencing method as the means to detect and identify SNPs in the genomic regions of human *BLM*, *WRN*, and *RECQL*. The aim was to develop a series of biallelic markers, 8–12 SNPs per gene, and use these to infer haplotypes that could allow chromosomal distinction. Although most efforts of defining variation in genes playing a role in cancer have concentrated on coding sequence (Helland et al. 1998; Josefsson et al. 1998; Storey et al. 1998; Li et al. 1999; Wagner et al. 1999), we focused on intronic and other noncoding sequences. A genomic sequence length of 150 kb was targeted so markers clustered within the region would potentially be in disequilibrium, facilitating the identification of functional variants of the gene through haplotype association.

The SNP data presented here allowed definition of haplotypes that mark the variation observed in the helicases in the four ethnic groups. Creation of complex haplotypes based on multiple markers can strengthen the power of association studies, and such data will be further useful in comparisons of haplotype frequencies and distributions across the human genome and the human population.

Table 1. Sequence Variants Identified in *BLM, WRN,* and *RECQL*.

Locus	SNP name	Position	Variation
BLM	**B4.1** (IVS1-2056lt-c)	21812[a]	T→C
	B4.2 (IVS1-20290g-a)	22083[a]	G→A
	B11 (IVS5-302a-g)	79526	A→G
	B16 (IVS16-954a-g)	118225	A→G
	B18.1 (IVS17-425a-g)	122762[a]	T→C
	B18.2 (IVS17-345c-g)	122842[a]	C→G
	B20 (IVS22-2082c-a)	136931[a]	G→T
	B21.1 (IVS22+3336c-g)	142615[a]	C→G
	B21.2 (IVS22+3401a-c)	142680[a]	A→C
	B22 (IVS22+9303c-t)	148582[a]	G→A
WRN	**W1** (IVS1-8213g-a)	6114[a]	G→A
	W5 (IVS4+176a-g)	45121[a]	T→C
	W10 (IVS17+1084c-a)	78453	C→A
	W12.1 (IVS19-3173t-c)	88968[a]	T→C
	W12.2 (IVS19-3145t-a)	88996[a]	T→A
	W15.1 (IVS24-210c-a)	111587	C→A
	W15.2 (IVS24-209a-g)	111588	A→G
	W15.3 (IVS24-191c-t)	111606[a]	G→A
	W18.1 (IVS32+845c-t)	135048[a]	G→A
	W18.2 (IVS32+859g-t)	135062[a]	G→T
	W20 (IVS34-628t-g)	145865[a]	A→C
	W22 (IVS35+4302t-c)	157465[a]	T→C
	W23 (IVS35+11737g-c)	164900[a]	G→C
	W26.1 (IVS53+30673c-t)	183836[a]	G→A
	W26.2 (IVS35+30764c-a)	183927[a]	G→T
RECQL	**R2** (IVS1-89964t-g)	10998[a]	A→C
	R8 (IVS1-55823a-g)	45139	A→G
	R10 (IVS1-4258lt-c)	58381[a]	T→C
	R12.1 (IVS1-30638g-c)	70324[a]	G→C
	R12.2 (IVS1-30329g-t)	70633[a]	G→T
	R13.1 (IVS1-24228g-a)	76734[a]	G→A
	R13.2 (IVS1-24159c-t)	76803[a]	G→A
	R15.1 (IVS1-7216a-g)	93746[a]	T→C
	R15.2 (IVS1-7166g-a)	93796[a]	G→A
	R19 (IVS10-1078g-a)	113771[a]	G→A
	R25 (IVS15+19546t-c)	152798[a]	T→C
	R27 (IVS15+33444t-c)	166696[a]	T→C

SNP positions base pairs are based on the reference sequence. In bold are SNP names used throughout this paper. B, W, and R are for *BLM, WRN,* and *RECQL,* respectively. Numbers denote the PCR product within which the SNP was detected in a 5′ to 3′ gene orientation. SNPs within the same PCR product are denoted ".1" and ".2." In parentheses are SNP names according to the Committee on Mutation Nomenclature (Adhoc Committee on Mutation Nomenclature, 1996). [a] denote SNPs used in the haplotype analysis.

RESULTS

SNP Detection and Distribution of Variation

The available genomic sequence spanned a region of ~154 kb, 186 kb, and 180 kb for *BLM*, *WRN*, and *RECQL*, respectively. Repetitive elements (40%–50% of each locus) were masked, and PCR primers were designed to generate products in the range of 350 bp to 600 bp for SNP detection. Products allowed scanning of between 15% and 20% of each locus by resequencing in five individuals.

The majority of the amplicons screened for polymorphisms spanned intronic or intragenic sequence (22/26 in *BLM*, 21/26 in *WRN*, and 27/29 in *RECQL*). A total of 15, 10, and 12 biallelic polymorphisms were identified in *WRN*, *BLM*, and *RECQL*, respectively (Table 1). All were noncoding and involved a single base change, the majority being transitions (6/10 in *BLM*, 8/15 in *WRN*, and 8/12 in *RECQL*). Therefore, SNPs were found at an average (over the three helicases) frequency of 1 in every 1290 bp. Figure 1 shows a schematic of the gene structure and the identified SNPs in each locus.

Allele Specific Oligonucleotide Hybridization (ASO) Analysis and Allele Frequencies

Eight of the 10 SNPs identified within *BLM*, 12 of the 15 SNPs in *WRN*, and 11 of the 12 SNPs in *RECQL* (Table 1) proved to be robust in PCR amplification and allele-specific oligonucleotide hybridization and were used in subsequent studies. Table 2 shows the frequencies of the rarer allele for each of the 31 SNPs detected in the three genes for each of the populations tested. Both alleles were observed in each ethnic group for 27 of 31 SNPs, with rare allele frequencies ranging from 0.01 to 0.5. The remaining four SNPs (W23, W26.1, W26.2,

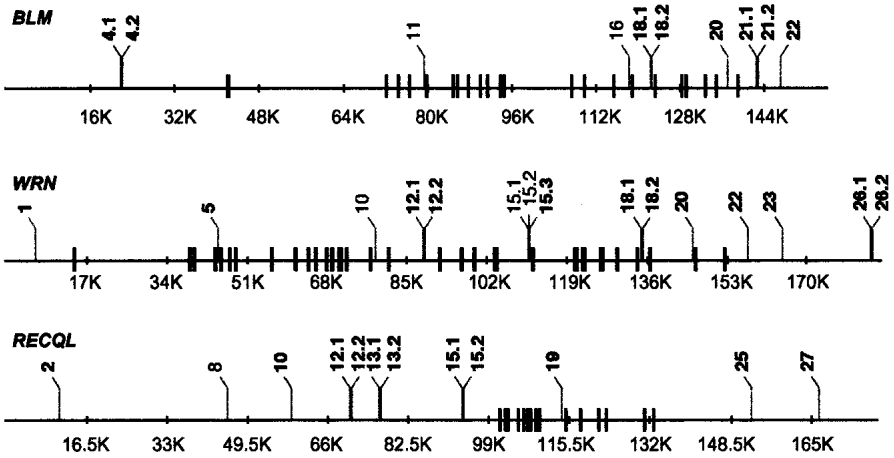

Figure 1 Schematic of SNP positions within *BLM*, *WRN*, and *RECQL*. Black vertical bars indicate exon positions and gray jagged lines the positions of identified SNPs in each gene. ".1" and ".2" denote SNPs found within the same PCR product. SNP numbering reflects their position in a 5′ to 3′ orientation. SNPs used in haplotype determination are in bold.

and R27) were excluded from the population analysis because they were monomorphic or contained very low frequency rare alleles.

Pairwise comparisons of allele frequencies between the different ethnic populations were carried out using the χ^2 contingency test. The majority of alleles had similar frequencies in all groups. Our data, however, showed the African American sample group to be the only one to differ significantly ($0.0001 < P < 0.001$) in at least two SNPs per gene. This finding may be related to the greater allelic diversity among individuals of African origin (Clark et al. 1998; Nickerson et al. 1998; Cargill et al. 1999; Halushka et al. 1999).

Hardy-Weinberg Equilibrium and Nucleotide Diversity

Tests for deviation from Hardy-Weinberg equilibrium (HWE) were conducted for each locus-population combination of SNP genotypes of the three helicases. Within *BLM*, the observed frequencies in the African American sample group deviated from the expected values at one locus (locus 22, $P \cong 0.0006$). Discrepancies were also observed at one locus of the *WRN* gene (W18.2) for the Caucasian population ($P \cong 0.0042$). In both cases a deficit of heterozygotes was observed. Given that 2 of 118 tests showed deviation from HWE, these significant variations can be ascribed to chance alone (*P*-values with the Bonferroni correction are 0.071 for B22 in African Americans and 0.496 for W18.2 in Caucasians). Also, no locus-specific trend of excess (or deficiency) of heterozygotes was found. Fisher's summed test statistic $-2\,\Sigma_i \ln p_i$ for the three helicases (summed over all locus-population tests) were equal to 66.79 ($P > 0.25$), 36.25 ($P > 0.995$), and 71.15 ($P > 0.50$) for BLM, RQL, and WRN, respectively (using the approximation of the distribution of the test statistic by the χ^2 distribution with $2k$ degrees of freedom, where k is the number of tests carried out).

Nucleotide diversity was measured by calculating π, the average heterozygosity per site in two sequences chosen from a randomly mating population (Nei 1987) in each ethnic group in each gene. Estimates of nucleotide diversity revealed no differences between populations at the levels of sequences spanned by the SNP sites. Mean values ranged from 0.0002 in *BLM* and *WRN* to 0.00033 in *RECQL* and were similar to those observed by Halushka et al. (1999) in a study of candidate genes for blood pressure homeostasis.

Haplotype Inference

EMHAPFRE (Excoffier and Slatkin 1995) was used to infer haplotypes and haplotype frequencies in the four ethnic groups (Caucasians, African American, Hispanics, and Asians) and in the members of the CEPH pedigrees. Taking into account inheritance patterns and assuming no recombination between generations, CEPH haplotypes were also determined manually and used as a control to test the accuracy of the EMHAPFRE algorithm. For the majority of the CEPH haplotypes, the two inference methods agreed. Small discrepancies were found among haplotypes of low frequencies. This is to be expected since EMHAPFRE is based on likelihood maximization and allows free recombination. Table 3 summarizes the results of the haplotype analysis in *BLM*, *WRN*, and *RECQL*.

The expectation-maximization (EM) algorithm predicted 50, 56, and 47 different haplotypes in 598, 618, and 616 chromosomes in *BLM*, *WRN*, and *RECQL*, respectively. Figure 2 shows the haplotype distribution for each gene in the four different ethnic groups based on their frequencies. African American haplotypes are more evenly distributed, reflecting the diversity observed in that population. Approximately 10% of haplotypes for each gene were found shared among all four ethnic groups. The shared haplotypes were also those with the highest frequencies, ranging from 0.02 to 0.3, as estimated by EMHAPFRE when all samples were studied together. These shared haplotypes also account for the majority (54%–94%) of the Caucasian, Hispanic, and Asian chromosomes at the three loci. In African Americans, however, a smaller percentage of samples (36%–53%) harbored the same shared haplotypes. In Figure 3, mean haplotype frequencies and number of haplotypes are plotted based on whether these are shared by one, two, three, or all the ethnic groups studied here. A small number of haplotypes with very low frequencies (total frequency of 0.009 in BLM and 0.018 in WRN and RECQL) appear in the EMHAPFRE output when all samples are pooled together. These are not present in the individual ethnic lists of haplotypes and are due to the larger data set analyzed by the program. Their impact was not considered significant and thus these haplotypes were not included in the graphs. As can be seen, the number of haplotypes shared by all populations is inversely related to their mean frequency. In other words, the majority of haplotypes are not shared by all four ethnic groups but have low frequencies, ranging from 0.009 to 0.04.

Linkage Disequilibrium and Recombination

Linkage disequilibrium (LD) between polymorphic sites in each gene was assessed in each population separately. We used haplotypic data for each ethnic group as determined by EMHAPFRE to calculate the *D*-value (Lewontin and Kojima 1960) for each pairwise comparison in DnaSP. Figure 4 shows

Table 2. Rare Allele Frequencies in *BLM, WRN,* and *RECQL*

BLM SNPs									
Population	N	B4.1 T/C	B4.2 G/A	B18.1 T/C	B18.2 C/G	B20 G/T	B21.1 C/G	B21.2 A/C	B22 G/A
AfAm	148	C.27	A.02	T.40	G.18	T.19	G.36	C.31	A.18
Asian	78	T.49	A.28	C.26	G.18	T.15	G.21	C.19	A.19
Cauc	154	C.28	A.12	C.29	G.20	T.30	G.37	C.39	A.15
Hisp	140	C.25	A.13	C.28	G.13	T.23	G.28	C.26	A.12
CEPH	78	C.27	A.12	C.36	G.26	T.26	G.33	C.33	A.24

WRN SNPs													
Population	N	W1 G/A	W5 T/C	W12.1 T/C	W12.2 T/A	W15.3 G/A	W18.1 G/A	W18.2 G/T	W20 A/C	W22 T/C	W23 G/C	W26.1 G/A	W26.2 G/T
AfAm	154	G.02	C.01	C.17	A.04	A.12	A.27	G.34	C.23	T.44	C.00	A.00	T.01
Asian	78	G.12	C.05	C.13	A.08	A.24	A.32	T.45	C.23	T.30	C.00	A.00	T.00
Cauc	158	G.04	C.06	C.22	A.20	A.27	A.30	G.42	C.25	T.47	C.03	A.03	T.01
Hisp	150	G.15	C.01	C.15	A.14	A.29	A.31	G.50	C.31	T.43	C.01	A.01	T.01
CEPH	78	G.01	C.04	C.21	A.21	A.21	A.31	G.44	C.37	T.43	C.01	A.01	T.01

RECQL SNPs												
Population	N	R2 A/C	R10 T/C	R12.1 G/C	R12.2 G/T	R13.1 G/A	R.13.2 G/A	R15.1 T/C	R15.2 G/A	R19 G/A	R25 T/C	R27 T/C
AfAm	156	C.10	C.38	G.47	G.47	G.47	A.19	C.19	G.49	A.30	C.47	C.01
Asian	74	C.01	C.30	C.46	T.46	A.45	G.45	T.45	A.45	G.43	C.05	C.00
Cauc	156	C.07	C.33	C.47	T.46	A.46	A.48	A.48	A.46	C.48	C.16	C.04
Hisp	152	C.03	C.34	C.47	T.46	A.46	A.45	A.46	A.42	G.43	C.17	C.01
CEPH	70	C.06	C.27	C.40	T.40	A.39	A.49	A.47	A.40	G.50	C.09	C.04

N denotes the number of chromosomes screened in each group. Only the grandparental chromosomes were considered in the CEPH pedigrees. In each case, the case of the rare allele is followed by its frequency. SNP names are as in Table 1.

the results of pairwise tests, indicating the site pairs with a significant Fisher's exact test ($P < 0.001$). Site pairs with rare alleles failed to give a significant Fisher's exact test outcome, although disequilibrium may exist. Analyses of individual populations excluded disequilibria computation for site pairs that include sites monomorphic in individual populations but polymorphic in the total sample.

As expected, SNPs found within the same PCR product showed significant disequilibrium. For *RECQL*, all populations showed LD throughout the gene, with the exception of the outer-most markers. Similarly, in *WRN* all ethnic groups showed comparable patterns of disequilibrium in the middle section of the gene (between the fifth and eighth SNPs). In *BLM*, however, disequilibrium was not confined to a specific gene region. The significance plots of LD (Fig. 4) were different in each population and only the Hispanic and Asian sample groups shared a similar pattern. In fact, these plots nearly mimic the plot of the absolute value of the normalized LD, D', for all three helicases in the four samples studied. In particular, the regions of significant LD (dark areas of Fig. 4) coincide with the regions of $D' = \pm 1$. More quantitatively, the proportion of pairwise $D' = \pm 1$, is the lowest for *BLM* (17.9%–28.6%) and the highest for *WRN* (60.6%–73.3%) across the four samples.

Despite the similarities of regions of pairwise LD, LD values (magnitude, as well as sign) are not the same in all populations. One explanation is that different demographic histories of the populations may have erased the signature of recombinations at different sites that occurred prior to their separation. Inferences of recombination events were made based on the commonly used four-gamete test (Hudson and Kaplan 1985) and the DnaSP recombination module using EMHAPFRE-predicted haplotypes. Intragenic recombination was inferred in all three genes. Figure 5 is a plot showing the four-gamete test results of pairwise comparisons, indicating in orange the site pairs where all four possible gametic phases were present. The asterisk denotes intervals when historical recombination events are suggested to have taken place. As expected, their locations are highly correlated with the ones implied by the four-gamete test.

Table 3. EMHAPFRE Inferred Haplotypes for *BLM, WRN,* and *RECQL*

	BLM		*WRN*		*RECQL*	
	H	N	H	N	H	N
CEPH	16(19)	78(194)	20(21)	78(194)	15(16)	68(178)
Af-American	22	146	23	154	21	156
Asian	16	78	14	78	10	74
Caucasian	25	152	29	158	20	156
Hispanic	22	144	21	150	27	152
Combined	50	598	56	618	47	616
Shared Haplotypes	5		4		6	

CEPH haplotype numbers inferred manually are in brackets; CEPH grandparents were used for EMHAPFRE prediction. The combined haplotype number was obtained when all ethnic samples and the CEPH grandparents were analyzed by EMHAPFRE. *H* denotes haplotypes and *N* the number of chromosomes used in each sample group. The number of CEPH chromosomes in parentheses reflects all family members to distinguish from the number present in the grandparents.

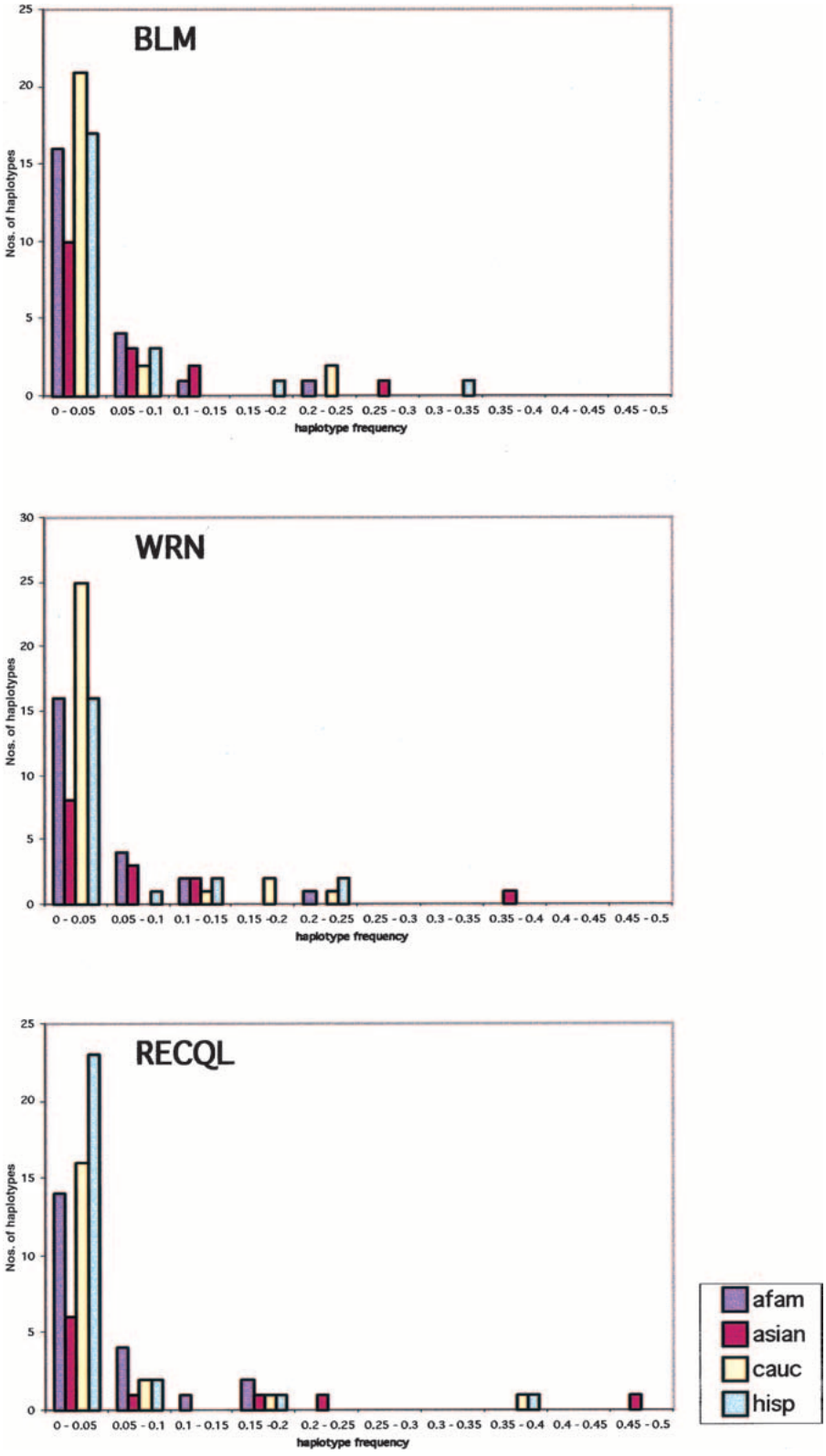

Figure 2 Haplotype distribution in the four ethnic groups for *BLM*, *WRN*, and *RECQL*. Haplotype numbers and their frequencies are plotted by ethnic background for the three genes. Frequencies are binned by .05 intervals. Afam, African American; asian, Asian; cauc, Caucasian; hisp, Hispanic.

Within *RECQL*, recombination is practically absent in the middle region of the gene consistent with the increased linkage disequilibrium observed in the same area. In *WRN*, however, the region of high linkage disequilibrium coincides with that of recombination, suggesting that the recombination events are probably of a recent origin. In *BLM*, recombination seems to have occurred throughout the gene and population patterns are not significantly different. In contrast, in *WRN* and *RECQL* Caucasians and Hispanics exhibit more recombination events than African Americans. This is an interesting observation because the African haplotype pool is expected to be older and, therefore, African Americans should exhibit increased numbers of recombinants compared to the other ethnic populations (Kidd et al. 1998, 2000). Ascertainment bias could be one factor influencing this observation (Hispanics have 60%–70% of their genes of European origin; Cerda-Flores et al. 1992). However, because *BLM* does not exhibit this pattern, sampling bias of the SNPs cannot entirely explain this finding.

Primate Sequence Comparisons and Phylogenetic Tree Construction

To assess the age of the detected polymorphisms, we screened five primate samples (two chimpanzees, one bonobo, and two gorillas), both by ASO hybridization and by direct sequencing. In all cases, the primate sites were homozygous for only one of the human alleles (Table 4). The primate nucleotide coincided with the more common human allele in 24 of 31 sites tested. These findings suggest that the observed human variation occurred after the separation of the ancestors of modern humans and great apes and that for the majority of the polymorphisms studied, the common human allele is likely to represent the state of sequence in the last common ancestor. All primates shared one haplotype per gene for the human-specific SNPs. The primate haplotypes were also found in the human haplotype pool, albeit at low frequencies (Table 4).

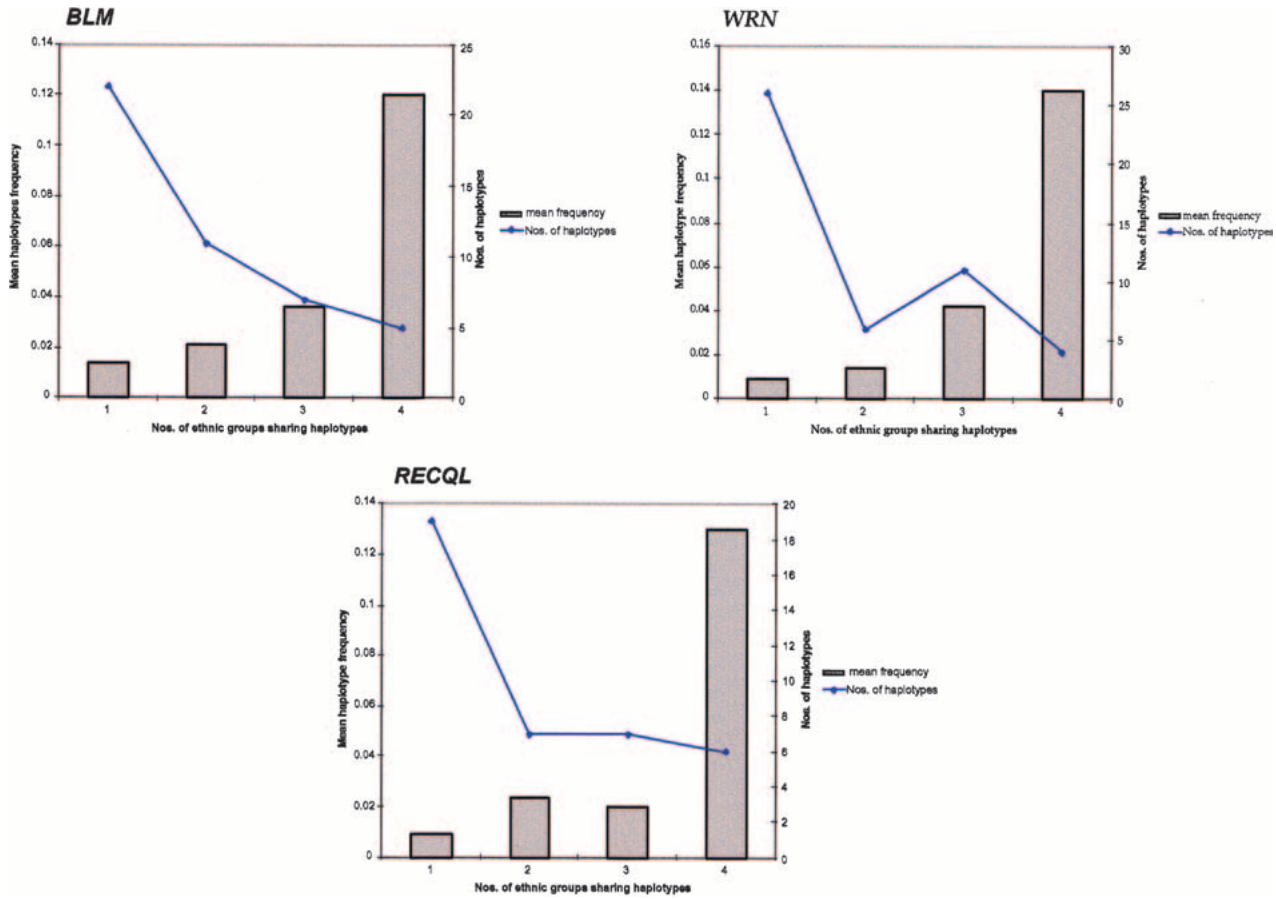

Figure 3 Ethnic sharing of haplotypes and their frequency for *BLM*, *WRN*, and *RECQL*. Haplotypes and their frequencies are plotted based on whether they are shared between one, two, three, or all four ethnic populations (*X*-axis). Bars indicate the mean frequencies of haplotypes shared; the line plots the number of haplotypes within each class.

In the phylogenetic analysis of the haplotypes, we assumed the primate haplotype to be the ancestral one for each gene and used it as the outgroup (data not shown). The relatively large number of haplotypes for all three genes made it difficult to define haplo groups. *RECQL* haplotypes can be divided into two major clades, but the *BLM* and *WRN* trees are bushlike. Assessment of the haplotype clusters based on whether they were part of the African American haplotype pool did not result in any particular pattern (data not shown). It is noteworthy that the internal branch lengths in these trees are generally small. Hence, any single evolutionary factor (e.g., recent population expansion, early history of recombination, or natural selection) is not sufficient explanation because the topologies differ from gene to gene.

Pilot Association Studies

The ultimate goal of this line of experiments is to create resources for association and predisposition studies. To test the utility of the identified SNPs and predicted haplotypes, we carried out two pilot association studies. The *blm*^{Ash} mutation (Ellis et al. 1995) is found in ~97% of Bloom syndrome patients of Ashkenazi Jewish origin, patients of Spanish American ancestry (Ellis et al. 1998), and Jewish patients of Polish descent (Shahrabani-Gargir et al. 1998). The mutation has a carrier frequency of ~1% among Ashkenazi Jews (Roa et al. 1999) and results in protein truncation due to 6-bp deletion/7-bp insertion in exon 10 (Ellis et al. 1995).

We tested four DNA samples and five cell lines from the Bloom syndrome patient registry, available from the Coriell Cell Repository. Of these, six are of Ashkenazic ancestry, two are listed as Caucasians, and one is of Japanese origin. All samples were tested for the *blm*^{Ash} mutation and the noncoding SNPs presented here by ASO analysis. As expected, all Ashkenazi Jewish samples were homozygous for the mutation, one Caucasian sample was heterozygous, and the remaining samples (one Caucasian and one Japanese) were negative. When we then studied haplotypes based on the *BLM* SNPs, we found the Ashkenazi Jews to be homozygous and share one haplotype (No. 25), which was also only present in a heterozygous state in the Caucasian sample heterozygous for the mutation. Based on the EMHAPFRE analysis, haplotype No. 25 is not seen in African Americans and its frequency ranges from 0.006 in Asians to 0.042 in Caucasians. When control samples of Caucasian origin carrying haplotype No. 25 in a heterozygous or homozygous state were tested for the *blm*^{Ash} mutation, they were found to be negative, showing that this haplotype predated the *blm*^{Ash} mutation. Our results are in agreement with previous haplotype studies (Ellis et al. 1998) that showed the presence of a predominant single haplotype (based on microsatellite markers surrounding *BLM*)

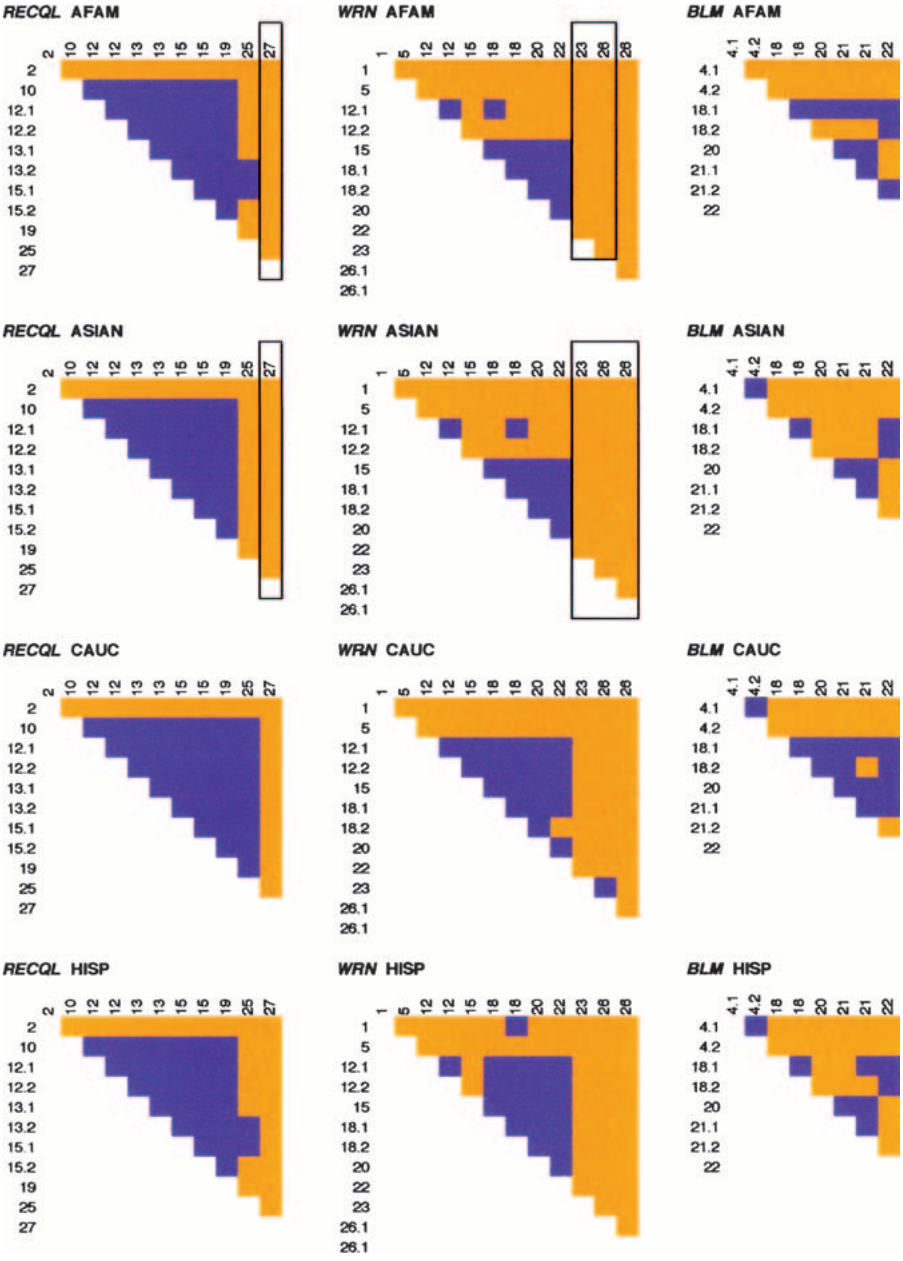

Figure 4 Linkage disequilibrium for each gene by ethnic group. Site pairs with significant Fisher's exact test (P <0.001) are shown in blue. Nonsignificant pairs are in orange. Loci are numbered in the 5′ to 3′; SNP order relative to the direction of transcription of each gene. ".1" and ".2" indicate SNPs found within the same PCR product. Framed loci were monomorphic and were not considered in the analysis. AFAM, African American; ASIAN, Asian; CAUC, Caucasian; HISP, Hispanic.

considered to be the founder haplotype that *blm*^{Ash} occurred in Ashkenazi Jews.

A second pilot association study utilized six previously described coding polymorphisms within the *WRN* gene (Meisslitzer et al. 1997; Ye et al. 1997; Vidal et al. 1998; Castro et al. 1999). Each of these was used separately in an attempt to establish association between any of the coding SNPs (cSNPs) and the *WRN* inferred haplotypes. A total of 136 individuals (40 African Americans, 40 Caucasians, 40 Hispanics, and 16 Asians) was assessed for noncoding SNP-based haplotypes and the presence or absence of coding variation. Allele frequencies were calculated for each cSNP in the four populations; the results are shown in Table 5.

Although allele frequencies differed among the four ethnic groups for some of the coding SNPs, no clear pattern could be observed between haplotype content and absence or presence of coding variation when all ethnic samples were pooled together. This is an apparent contradiction to our observation that the *WRN* haplotypes, in comparison to those in *BLM* gene, exhibit tighter LD (Fig. 4) and lesser recombination (Fig. 5). Recurrent mutations at the *WRN* cSNP sites may be one mechanism explaining the lack of association between cSNPs and noncoding SNPs at this gene. Two of the coding SNPs are in fact due to changes within CpG dinucleotides, supporting the possibility of recurrence.

Samples carrying coding SNPs cW4 or cW6 (c.2592T> *G* [Castro et al. 1999] and c.3453T> *G* [Meisslitzer et al. 1997], respectively) in a heterozygous or homozygous state for the low frequency allele (*G*) were studied based on their ethnic background. Hispanic individuals exhibited an increased frequency of haplotype No. 28 for both coding SNPs. This haplotype is shared among all ethnic groups showing the highest frequency in Hispanics (0.22 versus 0.16 in the pooled sample). A further increase of the frequency of haplotype No. 28 was noted among Hispanic individuals heterozygous and homozygous for the *G* allele at the cW4 (frequency 0.35) and cW6 (frequency 0.31) sites. These findings suggest an association between these polymorphisms and the chromosomal background of haplotype No. 28.

DISCUSSION

We chose SNPs to study variation because they are predominantly biallelic, abundant (1 in ~1000 bp), and stable (Wang et al. 1998). We concentrated our efforts to detect variations in intronic/intragenic regions as these are expected to be more variable, while avoiding repeat sequences that would complicate analysis. Particular care was taken to disperse the PCR products/sequencing targets throughout the genomic regions surrounding each gene. A target of 8–12 SNPs per gene was set to identify adequate numbers of loci evenly distributed along the gene region and thus create informative haplotypes that would enable differentiation between chromosomes. Coding

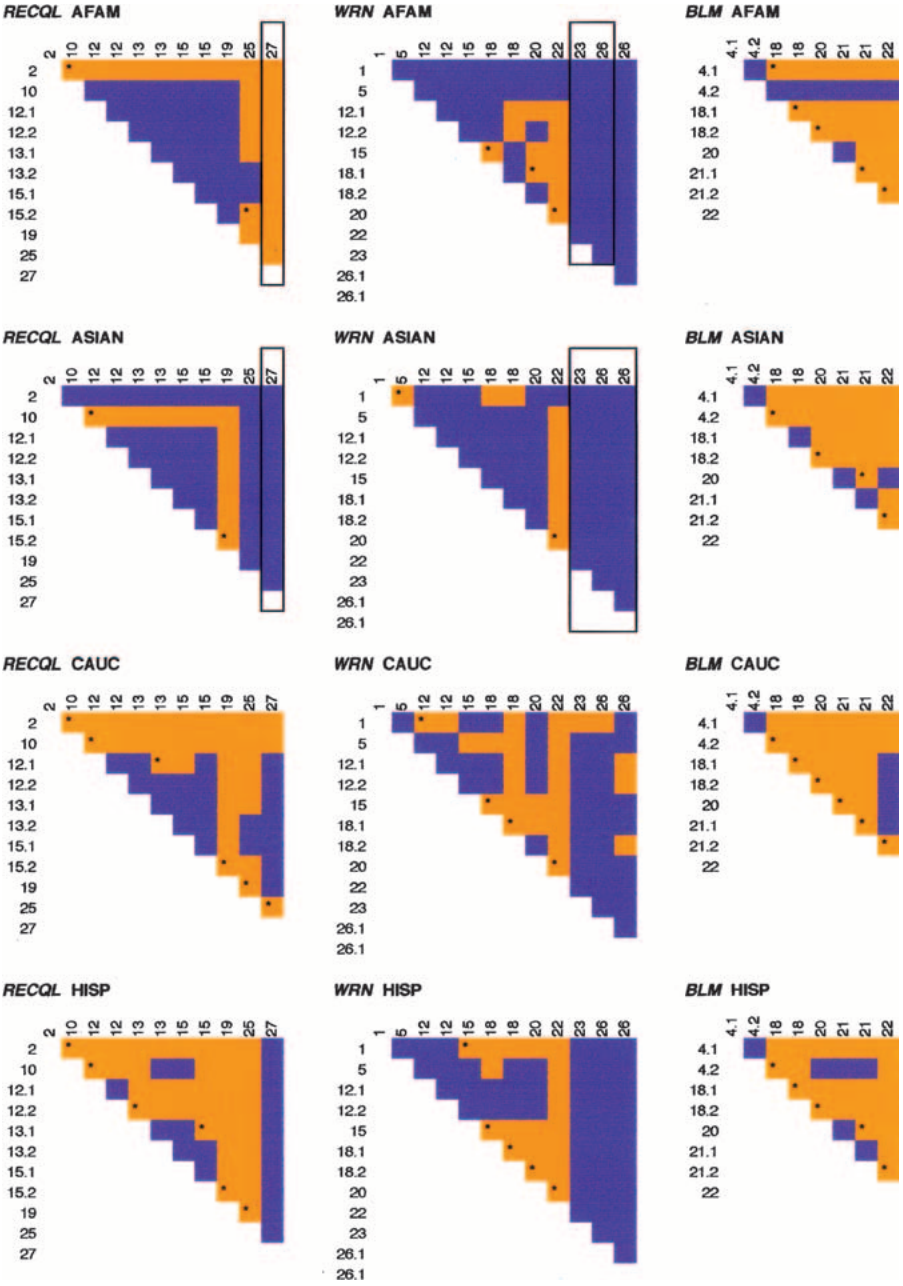

Figure 5 Recombination for each gene by ethnic group. Orange indicates site pairs positive for the four-gamete test; blue indicates absence of one or more possible gametes. Presence of all four gametic phases suggests historical intragenic recombination. The minimum number of recombination events is indicated by asterisks. Loci are numbered in the 5′ to 3′ SNP order relative to the direction of transcription of each gene. ".1" and ".2" indicate SNPs found within the same PCR product. Framed loci were monomorphic and were not considered in the analysis. AFAM, African American; ASIAN, Asian; CAUC, Caucasian; HISP, Hispanic.

sequences were generally avoided to reduce the likelihood of bias toward functional variants.

Variation in *BLM*, *WRN*, and *RECQL*

From the observation that each of the 31 SNP sites examined showed no variation in five primate samples (two chimpanzees, two gorillas, and one bonobo), we postulate that the ages of the polymorphism in human are probably younger than the time of separation of human and ape lineage. Because the systematic search for SNPs in the noncoding regions of these helicases was made from sequence analysis of a small number (10) of Caucasian chromosomes, it is clear that we detected a small fraction of all polymorphic sites in the gene regions. Nevertheless, there are at least three lines of evidence suggesting that our observations are not severely affected by the ascertainment bias due to selection of sites.

One type of ascertainment bias might stem from the fact that the sequenced chromosomes belong to Caucasians. To determine if such bias is present in the data, we carried out computations of sequence diversities (Nei 1987) of the SNPs at the *BLM*, *WRN*, and *RQL* genes separately for each of the populations considered. In addition, we computed haplotype diversites at these genes in all populations. In this latter set of computations, each haplotype was treated as a different allele, identifiable by its SNP sequence. Detailed numerical result are depicted in Supplementary Table 3 available online at http://www.genome.org. The trends observed are as follows.

Sequence diversities (Nei 1987), in most cases, assume highest values for Caucasian or CEPH individuals, whereas African Americans, Asian Americans, and Hispanics are characterized by systematically lower values of this index. Although no formal test was carried out, the differences exceed the three-standard-deviations level. On the contrary, the haplotype diversities are systematically higher for African Americans and lower for the other subpopulations, again with the differences exceeding the fluctuation level.

Precise interpretation of these findings requires more study. However, it seems that although the effects of ascertainment bias are felt at the level of individual SNPs (Nei's sequence diversity is equal to the average of individual SNP diversities although these two indices have different variances; Nei 1987), they are not detectable at the level of haplotypes. This seems to constitute an argument for using haplotype-based statistics as being more resistant to ascertainment bias.

Another type of ascertainment bias might stem from the

Table 4. Primate Haplotypes and Their Frequencies in Humans

Primate haplotype	African American	Asian	Caucasian	Hispanic	All samples
BLM CGCCGCAG	0.07	0.02	NP	NP	0.02
WRN GTTTGGGATGGG	0.016	NP	NP	0.021	0.001
RECQL ATGGAGTGGTT	0.006	NP	NP	NP	0.002

Frequencies in each ethnic group are those calculated by EMHAPFRE.
NP, not predicted haplotypes and haplotypic frequencies.

small number (10) of sequenced chromosomes, independently of the population from which they were taken. This type of bias was recently analyzed by us (A. Renwick et al., in prep.) and indicates that SNP data, collected with a screening protocol such as ours, results in a small (<10%) underestimation of nucleotide diversity.

Even though we did not detect the presence of more than two alleles in any of the SNP sites, we realize that individuals carrying a rare third or fourth allele would have been mistyped as homozygotes with our ASO-based genotyping assay. However, two features of our data rule out this as a major factor. First, the primate sequences revealed cross-species monomorphism for one of the human variants at each polymorphic site. Second, the statistical agreement of the observed and expected genotype frequencies (under HWE) also argues for the absence of a third or fourth allele. As mentioned before, the two departures from HWE expectation (B22 in African Americans and W18.2 in Caucasians,) can be ascribed to chance alone by Bonferroni adjustment of multiple testing, as well as Fisher's summed statistic of individual *P*-values. Further, there is no site-specific trend of heterozygote deficiencies, also suggestive of no allele drop-out effect of the ASO-based genotype assay.

Haplotype Prediction

The algorithm used (EMHAPFRE) for deconvoluting haplotypes from genotype data assumes that the genotype frequencies at every SNP site follow the expectations of Hardy-Weinberg equilibrium and that there is free recombination between sites. Although the first assumption appears approximately valid for our data, there is no direct validation for the second one. In fact, manual checking of the genotype data from the CEPH family samples showed a somewhat larger number of haplotypes (19 for *BLM*, 21 for *WRN*, and 16 for *RECQL*, instead of 16, 20, and 15, respectively, inferred by EMHAPFRE). The extra haplotypes detected through manual inference, however, are low frequency ones. We also verified the accuracy of haplotype inference by using a newly developed data-mining algorithm (N. Wang et al., in prep.) that depends on neither the assumption of HWE nor of free recombination. In general, common haplotypes inferred by the EM algorithm are also detected by the data-mining algorithm. The assumption of free recombination in EMHAPFRE results in a somewhat larger number of inferred haplotypes. The extra haplotypes are always rare. Hence, they contribute little to estimates of haplotype diversity and interpopulation variation. Despite these differences, the number of inferred haplotypes for each helicase is much larger than the number of polymorphic sites examined (50 haplotypes for the 8 *BLM* sites, 56 for the 12 *WRN* sites, and 47 for the 11 *RECQL* sites). This clearly indicates that recombination and/or recurrent mutation have been active in producing these haplotypes because unique mutations at these sites can only produce ($s + 1$) haplotypes with s polymorphic sites (Fu and Li 1993). From genotype data, we ruled out recurrent mutations producing more than two nucleotides as a major factor. However, the possibility of back mutation cannot be distinguished from recombination in the haplotype inference method. The number of segregating sites discovered is different among the helicases. Furthermore, the extent of haplotype variation, expressed in terms of number of haplotypes (Table 3), as well as haplotype frequencies (Fig. 2), differ among these loci.

Recombination Events and Linkage Disequilibrium

Despite the fact that we used a smaller number of segregating sites (8) for the *BLM* gene in comparison to the others, the resulting number of haplotypes for this helicase is nearly the same as for the others. Inference of recombination events (by the four-gamete test) and plotting significant LD values by pairwise location of sites (Figs. 4, 5) provide an explanation for this observation. In *BLM*, recombination appears to have occurred throughout the gene in each of the four ethnic groups. In *RECQL*, recombination events appear to occur predominantly at the 5′ and 3′ ends of the gene, leaving a middle section free of inferred recombination. The opposite (i.e., recombinations occurring in the center of the gene) is seen in *WRN*. This is consistent with the pattern of LD (Fig. 4) among the sites for these genes as well. For *BLM*, the tendency of more significant LD from the third SNP-extending 3′ is consistent with the physical location of the sites (Fig. 1; Table 1) because the second and third SNP are ~100 kb apart.

The recombination patterns and linkage disequilibria observed in the African American sample are somewhat unexpected (Figs. 4, 5). For *WRN* and *RECQL*, the African American and the Asian samples exhibit similar patterns of recombination, which appear to be less when compared to the Caucasian and the Hispanic samples. Although not studied here, Africans are expected to have greater diversity reflecting their antiquity and, hence, recombinations should have been more prevalent in them. In all three genes, the Caucasian group shows the highest number of recombination events and haplotypes, with the exception of *RECQL*, in which the Hispanic sample has the maximum number of haplotypes. The African American and the Hispanic samples are populations of admixed origin, whereas the Caucasian sample is of primarily European ancestry. Therefore, these paradoxical observations cannot be readily explained by gene-specific recombination alone. Whether ascertainment bias (i.e., the SNP-screening protocol) or recurrent mutations are responsible should be studied in more detail.

Our observation that significant LD is found for all three genes in the four ethnic groups within 15 kb in *BLM*, ~70 kb in *WRN*, and 100 kb in *RECQL* are relevant for the general utility of LD-based gene-mapping approaches. Recent studies suggest a range of 3 kb (Kruglyak 1999) to 500 kb (Thompson and Neel 1997) as target regions within which LD-based map-

Table 5. Allele Frequencies of *SRN*-Coding SNPs

	cW1		cW3		cW4		cW6		cW7		cW8	
	T	C	G	A	T	G	T	G	C	T	T	C
AfAm (*N* = 80)	0.54	0.46	0.97	0.03	0.51	0.49	0.64	0.36	0.85	0.15	0.86	0.14
Asian (*N* = 32)	0.8	0.2	1	0	0.54	0.46	0.5	0.5	0.82	0.18	0.87	0.13
Caucasian (*N* = 80)	0.72	0.28	0.91	0.09	0.47	0.53	0.47	0.53	0.77	0.23	0.69	0.31
Hispanic (*N* = 80)	0.75	0.25	0.91	0.09	0.54	0.46	0.54	0.46	0.72	0.28	0.82	0.18

N = number of chromosomes.
cW1 = c.744T > C[a], cW3 = c.1392G > A[b,c], cW4 = c.2592T > G[a], cW6 = c.3453T > G[b], cW7 = c.4314C > T[a], cW8 = c.4330T > C[d].
[a]Castro et al. 1999.
[b]Meisslitzer et al. 1997.
[c]Vidal et al. 1998.
[d]Ye et al. 1997.

ping approaches is feasible. Our data falls within this range, but the pattern of LD appears to be gene specific. Therefore, we argue that the genomic context should be considered as a factor in LD-based mapping strategies. We showed that the number, as well as the density of SNP markers that can produce similar haplotype diversity, can vary among loci. In a study similar to ours, Bonnen et al. (2000) showed that at the *ATM* gene region, a longer genomic region (142 kb) carries more restricted recombination and, consequently, tighter LD.

Pilot Association Studies

Our pilot association studies support the utility of SNP-based haplotypes in identifying mutations of disease relevance. In the first pilot study, we identified a haplotype (No. 25) in the *BLM* gene, which appears to be the ancestral haplotype on which the *blm*^{Ash} mutation occurred in the Ashkenazi Jewish population. Ellis et al. (1998) using microsatellite data obtained similar association results. However, the stability of SNP sites offers opportunities to estimate the age of such disease mutations more precisely.

The importance of ethnically matched controls in disease-gene association studies is revealed in the data of our second pilot study. In particular, the increased frequency of *WRN* haplotype No. 28 (0.35 compared to 0.22) seen in the Hispanic individuals who are either homozygous or heterozygous for one or both of the two cSNP mutations (c.2592T > G and c.3453T > G) was no longer found in the pooled sample of the four ethnic groups.

Primate Variation in the Respective Loci

Studying the sequence variation in five primate samples at these gene regions served two purposes. First, we observed no variation at any of the SNP sites for all three helicases in the primates. Further, for all three genes, the primate haplotypes were indeed observed in the human populations, albeit in low frequencies (0.02, 0.001, and 0.002 for *BLM*, *WRN*, and *RECQL*, respectively). These helped us infer the ancestral haplotypes in human, allowing rooting of the haplotype phylogeny in our tree reconstruction effort. Second, based on the approximate time of divergence of chimpanzees and human (5 million years, or ~200,000 generations) and the average frequency of SNP occurrence per base-pair distance (1 in 1 kb), we can estimate the average rate of nucleotide substitution in these three genes. The result (5×10^{-9} per base pair per generation) coincides with the generally accepted value (Nei 1987).

In conclusion, our data show that even with nearly equal nucleotide diversity in different gene regions, haplotype diversity based on several SNP sites may have a very different pattern across genes. Differences of demographic history of populations, superimposed on different patterns of recombination among genes, can result in different haplotype distributions across populations. Therefore, careful ethnic matching and/or family-based sampling designs such as the TDT (Spielman et al. 1993) should be a critical aspect of SNP-based disease-gene association studies.

METHODS

DNA Samples

Lymphoblastoid cell line samples from five unrelated Caucasians were used in screening the genomic regions of *BLM*, *WRN*, and *RECQL* for single nucleotide polymorphisms. ASO analysis was carried out in ten three-generation CEPH families (parents, four children, and paternal and maternal grandparents) and samples from four ethnic groups: Caucasians, African Americans, Hispanics, and Asians. The ethnic samples were a kind gift from Drs. Michael Weil and Michael Story (MD Anderson, Houston, TX). Random, ethnically self-described blood donors in Houston were asked to contribute under informed consent. Blood samples were anonymized and DNAs were coded with ethnic identifiers. A minimum of 70 individuals was screened for each gene, with the exception of the Asian group, which had only 39 samples available. DNA samples and cell lines used in the association studies for the *blm*^{Ash} mutation were obtained from Coriell Cell Repositories (GM00811, GM01492, GM03510, GM02085, GM02520, NA04408, NA09960, NA03403, and NA05289). Of these, six are from individuals of Ashkenazi Jewish origin, homozygous for the mutation. Primate genotypes were ascertained by screening two chimpanzee (*Pan troglodytes*, PTR), one bonobo (*Pan paniscus*, PPA), and two gorilla (*Gorilla gorilla*, GGO) cell lines for the human SNPs. Primate cell lines were kindly provided by Dr. E. Nickerson, Cold Spring Harbor Laboratories.

Analysis of Genomic Sequences

BLM and *WRN* genomic sequences were obtained from Gen-

Bank (accession nos. AC002312 for *BLM* and AF181896 and AF181897 for *WRN*). Complete genomic sequence for the RECQL locus was developed in the course of this study from a bacterial artificial chromosome clone from a human library (RPCI11–501E26) and deposited into GenBank under accession no. AC006559.

In all three genes, the genomic sequence was masked for known repetitive elements with RepeatMasker (RepeatMasker available at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The remaining unmasked sequence was used with the MacVector software package to design PCR primers that would amplify single-copy regions of the genes. Exons were avoided where possible.

SNP Identification and Detection

PCR products from five unrelated Caucasians (10 chromosomes) were sequenced using ^{33}P terminators (ThermoSequenase Radiolabeled Terminator Cycle Sequencing kit, USB Pharmaceuticals). Reaction products were visualized by autoradiography of dried polyacrylamide gels. Polymorphic PCR products were combined into multiplex PCR groups for the screening of the ethnic samples. The PCR conditions and the primer sequence for the multiplex groups for the three genes are listed in Supplementary Table 1 available online at http://www.genome.org. Only the conditions for the SNPs used in the haplotype analysis (see Results) are given.

PCR reactions were carried out in a GeneAmp PCR System 9700 Thermocycler (Perkin Elmer) for 35 cycles at 50 µL final volume with 1× reaction buffer (10× PCR Buffer, Perkin Elmer), 250 µM each dNTP (Amersham), 0.5U Taq (Perkin Elmer), and 1 µM each primer, except where otherwise stated. In some reactions 4%–8% DMSO was added to facilitate the annealing of the primer to the template. Six known coding polymorphisms within the *WRN* gene (Meisslitzer et al. 1997; Ye et al. 1997; Vidal et al. 1998; Castro et al. 1999) were used in an association pilot study. PCR products spanning the polymorphisms were pooled together into two multiplex groups (Supplementary Table 1 available online at http://www.genome.org.) and amplified in 40 Caucasians, 40 African Americans, 40 Hispanics, and 16 Asians. Numbering of coding SNPs reflects their order from 5′ to 3′ of the gene. Prior to sequencing, all PCR products were purified with the PCR product presequencing kit (Amersham). Unwanted remnants of PCR reactions were enzymatically removed using equal volumes of Shrimp alkaline phosphatase (200U) and Exonuclease I (1000U) and incubating samples at 37°C for 15 min and then at 80°C for an additional 15 min.

ASO Hybridization

The protocol that was followed for the ASO design, dot blot of PCR samples, and ASO hybridization has been previously described in DeMarchi and colleagues (1994). Multiplexed PCR products were spotted onto GeneScreen Plus nylon membranes (New Life Science). All ASO probes were 19mers with the variable site usually at the tenth position. As suggested by DeMarchi et al. (1994), to avoid false-positives, the DNA strand from which the ASO was chosen was the one that avoided G:T and G:A mismatches between mutant oligonucleotide and wild-type DNA template. SNPs that did not result in robust PCR products or hybridization signals were not included in the ASO analysis. A total of 16, 24, and 22 ASOs (two oligonucleotides per SNP) were designed and used in the analysis for *BLM*, *WRN*, and *RECQL*, respectively (Supplementary Table 2 available online at http://www.genome.org.).

Forty picomoles of oligonucleotide were 5′-end labeled with γ-^{32}P ATP and purified through G25 Sephadex Quick Spin columns (Boehringer) each time. Labeling reactions were at 100 µL final volume. PCR products were spotted onto membranes which were prehybridized in 5 mL hybridization buffer (DeMarchi et al. 1994) at 65°C for at least 30 min. Hybridization was carried out in the same buffer after adding 25 µg/mL sheared salmon sperm DNA, 80 pmoles the alternative nonradioactive oligo, and one-fifth labeling reaction. Hybridization bags containing the membranes were submerged in a water bath for 30 min at 65°C. The temperature was then adjusted to 34°C and the bags were left to cool overnight with shaking. Membranes were then washed collectively at room temperature in 5× SSC for 5 min twice. Background was removed with an additional wash in 2× SSC for 30 min at a temperature that was empirically determined for each probe. Membranes were exposed to X-omat AR film (Kodak) overnight.

Screening of the samples for the coding SNPs was carried out by ASO hybridization as for the noncoding SNPs. ASO sequences and wash conditions for each gene are shown in Supplementary Table 2 available online at http://www.genome.org. Amplification of the region spanning the *blm*^{Ash} mutation and screening of patients by ASOs were carried out as described by Roa et al (1999).

Haplotype Inference

Haplotypes were ascertained using EMHAPFRE, a computer-assisted statistical analysis based on maximum likelihood. EMHAPFRE infers haplotypes and haplotype frequencies under the assumption of Hardy-Weinberg equilibrium and implementing an expectation-maximization (EM) algorithm. For the CEPH pedigrees, haplotypes were also assigned by inspection as follows. Once a complete homozygote was identified within a family, a haplotype was unambiguously assigned. If an individual had a single heterozygous site, then two haplotypes were unambiguously identified. The remaining unresolved sequences were determined based on the known haplotypes and the inheritance pattern under the assumption of zero recombination events between generations. This method is easily applicable to related samples and the haplotypes of the 10 families studied could be unambiguously assigned. The four ethnic groups (Caucasians, African American, Hispanics, and Asians), however, comprised unrelated individuals and therefore the deduced haplotypes were the result of the EMHAPFRE-based analysis.

Assessing Nucleotide Diversity, Linkage Disequilibrium, and Recombination

DnaSP version 3.14 software package (Rozas and Rozas 1995, 1999) was used to estimate nucleotide variation and test for linkage disequilibrium (LD) and recombination between polymorphic sites in each gene locus. The measure of sequence variability that was calculated in this study is π (Nei 1987), the average number of differences per site between two sequences. The degree of LD is estimated by computing D (Lewontin and Kojima 1960) and D′ (Lewontin 1964), and Fisher's exact test was used to determine significance of associations between polymorphic sites. Positions for recombination events were determined using the recombination module of DnaSP. This module calculates the minimum number of recombination events based on the four-gamete test, as described by Hudson and Kaplan (1985).

ACKNOWLEDGMENTS

We are indebted to Drs. Michael Weil and Michael Story for providing the ethnic samples and Dr. Elizabeth Nickerson for providing the primate cell lines. We also thank Dr. Ning Wang for providing the Data Mining software package and Dr. Nessan A. Bermingham for constructive comments throughout the preparation of this paper. This work was supported by a grant from the National Cancer Institute of the NIH (CA75432).

The publication costs of this article were defrayed in part

by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adhoc Committee on Mutation Nomenclature. 1996. Update on human gene mutations. *Hum. Mutat.* **8:** 197–202.

Bonnen, P.E., Story, M.D., Ashorn, C.L., Buchholz, T.A., Weil, M.M., and Nelson, D.L. 2000. Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67:** 1437–1451.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231–238.

Castro, E., Ogburn, C.E., Hunt, K.E., Tilvis, R., Louhija, J., Penttinen, R., Erkkola, R., Panduro, A., Riestra, R., Piussan, C., et al. 1999. Polymorphisms at the Werner locus: I. Newly identified polymorphisms, ethnic variability of 1367Cys/Arg, and its stability in a population of Finnish centenarians. *Am. J. Med. Genet.* **82:** 399–403.

Cerda-Flores, R.M., Kshatriya, G.K., Berion, T.K., Hewett-Emmett, D., Hanis, C.L., and Chakraborty, R. 1992. Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. *Ann. Hum. Biol.* **19:** 347–360.

Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., et al. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63:** 595–612.

DeMarchi, J.M., Richards, C.S., Fenwick, R.G., Pace, R., and Beaudet, A.L. 1994. A robotics-assisted procedure for large scale cystic fibrosis mutation analysis. *Hum. Mutat.* **4:** 281–290.

Ellis, N.A., Groden, J., Ye, T.Z., Straughen, J., Lennon, D.J., Ciocci, S., Proytcheva, M., and German, J. 1995. The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83:** 655–666.

Ellis, N.A., Ciocci, S., Proytcheva, M., Lennon, D., Groden, J., and German, J. 1998. The Ashkenazic Jewish Bloom syndrome mutation *blm*^{Ash} is present in non-Jewish Americans of Spanish ancestry. *Am. J. Hum. Genet.* **63:** 1685–1693.

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

Foucault, F., Vaury, C., Barakat, A., Thibout, D., Planchon, P., Jaulin, C., Praz, F., and Amor-Gueret, M. 1997. Characterization of a new BLM mutation associated with a topoisomerase II α defect in a patient with Bloom's syndrome. *Hum. Mol. Genet.* **6:** 1427–1434.

Fu, Y. and Li, W.H. 1993. New statistical tests of neutrality for DNA samples from population. *Genetics* **133:** 693–709.

Fukuchi, K., Martin, G.M., and Monnat Jr., R.J. 1989. Mutator phenotype of Werner syndrome is characterized by extensive deletions. *Proc. Natl. Acad. Sci.* **86:** 5893–5897.

German, J. 1993. Bloom syndrome: A mendelian prototype of somatic mutational disease. *Medicine* **72:** 393–406.

Goddard, K.A., Hopkins, P.J., Hall, J.M., and Witte, J.S. 2000. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66:** 216–234.

Goto, M., Tanimoto, K., Horiuchi, Y., and Sasazuki, T. 1981. Family analysis of Werner's syndrome: A survey of 42 Japanese families with a review of the literature. *Clin. Genet.* **19:** 8–15.

Goto, M., Miller, R.W., Ishikawa, Y., and Sugano, H. 1996. Excess of rare cancers in Werner syndrome (adult progeria). *Cancer Epidemiol. Biomarkers Prev.* **5:** 239–246.

Gray, M.D., Shen, J.C., Kamath-Loeb, A.S., Blank, A., Sopher, B.L., Martin, G.M., Oshima, J., and Loeb, L.A. 1997. The Werner syndrome protein is a DNA helicase. *Nat. Genet.* **17:** 100–103.

Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22:** 239–247.

Helland, A., Langerod, A., Johnsen, H., Olsen, A.O., Skovlund, E., and Borresen-Dale, A.L. 1998. p53 polymorphism and risk of cervical cancer. *Nature* **396:** 530–531.

Hudson, R.R. and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111:** 147–164.

Josefsson, A.M., Magnusson, P.K., Ylitalo, N., Quarforth-Tubbin, P., Ponten, J., Adami, H.O., and Gyllensten, U.B. 1998. p53 polymorphism and risk of cervical cancer. *Nature* **396:** 531.

Karow, J.K., Chakraverty, R.K., and Hickson, I.D. 1997. The Bloom's syndrome gene product is a 3′–5′ DNA helicase. *J. Biol. Chem.* **272:** 30611–30614.

Kidd, K.K., Morar, B., Castiglione, C.M., Zhao, H., Pakstis, A.J., Speed, W.C., Bonne-Tamir, B., Lu, R.B., Goldman, D., Lee, C., et al. 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum. Genet.* **103:** 211–227.

Kidd, J.R., Pakstis, A.J., Zhao, H., Lu, R.B., Okonofua, F.E., Odunsi, A., Grigorenko, E., Tamir, B.B., Friedlaender, J., Schulz, L.O., et al. 2000. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.* **66:** 1882–1899.

Knudson, A.G. 1996. Hereditary cancer: Two hits revisited. *J. Cancer Res. Clin. Oncol.* **122:** 135–140.

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.

Kwok, P.Y., Deng, Q., Zakeri, H., Taylor, S.L., and Nickerson, D.A. 1996. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* **31:** 123–126.

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

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:** 49–67.

Lewontin, R.C. and Kojima, K. 1960. The evolutionary dynamics of complex polymorphisms. *Evolution* **14:** 450–472.

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

Long, A.D. and Langley, C.H. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9:** 720–731.

Martin, E.R., Gilbert, J.R., Lai, E.H., Riley, J., Rogala, A.R., Slotterbeck, B.D., Sipe, C.A., Grubber, J.M., Warren, L.L., Conneally, P.M., et al. 2000. Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* **63:** 7–12.

Meisslitzer, C., Ruppitsch, W., Weirich-Schwaiger, H., Weirich, H.G., Jabkowsky, J., Klein, G., Schweiger, M., and Hirsch-Kauffmann, M. 1997. Werner syndrome: Characterization of mutations in the WRN gene in an affected family. *Eur. J. Hum. Genet.* **5:** 364–370.

Moser, M.J., Oshima, J., and Monnat, Jr., R.J. 1999. WRN mutations in Werner syndrome. *Hum. Mutat.* **13:** 271–279.

Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.

Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., and Sing, C.F. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19:** 233–240.

Ober, C., Leavitt, S.A., Tsalenko, A., Howard, T.D., Hoki, D.M., Daniel, R., Newman, D.L., Wu, X., Parry, R., Lester, L.A., et al. 2000. Variation in the interleukin 4-receptor α gene confers susceptibility to asthma and atopy in ethnically diverse populations. *Am. J. Hum. Genet* . **66:** 517–526.

Puranam, K.L. and Blackshear, P.J. 1994. Cloning and characterization of *RECQL*, a potential human homologue of the *Escherichia coli* DNA helicase RecQ. *J. Biol. Chem.* **269:** 29838–29845.

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

Roa, B.B., Savino, C.V., and Richards, C.S. 1999. Ashkenazi Jewish population frequency of the Bloom syndrome gene 2281 delta 6ins7 mutation. *Genet. Test* **3:** 219–221.

Rozas, J. and Rozas, R. 1995. DnaSP, DNA sequence polymorphism: An interactive program for estimating population genetics parameters from DNA sequence data. *Comput. Appl. Biosci.* **11:** 621–625.

———. 1999. DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174–175.

Shahrabani-Gargir, L., Shomrat, R., Yaron, Y., Orr-Urtreger, A., Groden, J., and Legum, C. 1998. High frequency of a common Bloom syndrome Ashkenazi mutation among Jews of Polish origin. *Genet. Test* **2:** 293–296.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52:** 506–516.

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

Thompson, E.A. and Neel, J.V. 1997. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* **60:** 197–204.

Umezu, K., Nakayama, K., and Nakayama, H. 1990. *Escherichia coli* RecQ protein is a DNA helicase. *Proc. Natl. Acad. Sci.* **87:** 5363–5367.

Vidal, V., Bay, J.O., Champomier, F., Grancho, M., Beauville, L., Glowaczower, C., Lemery, D., Ferrara, M., and Bignon, Y.J. 1998. The 1396del A mutation and a missense mutation or a rare polymorphism of the *WRN* gene detected in a French Werner family with a severe phenotype and a case of an unusual vulvar cancer. *Hum. Mutat.* **11:** 413–414.

Wagner, T.M., Hirtenlehner, K., Shen, P., Moeslinger, R., Muhr, D., Fleischmann, E., Concin, H., Doeller, W., Haid, A., Lang, A.H., et al. 1999. Global sequence diversity of BRCA2: Analysis of 71 breast cancer families and 95 control individuals of worldwide populations. *Hum. Mol. Genet.* **8:** 413–423.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077–1082.

Ye, L., Miki, T., Nakura, J., Oshima, J., Kamino, K., Rakugi, H., Ikegami, H., Higaki, J., Edland, S.D., Martin, G.M., et al. 1997. Association of a polymorphic variant of the Werner helicase gene with myocardial infarction in a Japanese population. *Am. J. Med. Genet.* **68:** 494–498.

Yu, C.E., Oshima, J., Fu, Y.H., Wijsman, E.M., Hisama, F., Alisch, R., Matthews, S., Nakura, J., Miki, T., Ouais, S., et al. 1996. Positional cloning of the Werner's syndrome gene. *Science* **272:** 258–262.

Received July 30, 2001; accepted in revised form January 29, 2002.