Genetic Diversity Patterns in the SR-BI/II Locus Can Be Explained by a Recent Selective Sweep

Mireille Le Jossec,* Tina Wambach,* Damian Labuda,** Daniel Sinnett,**† and Emile Levy**†

*Centre de Recherche, Hôpital Sainte-Justine, Montréal, Quebec, Canada; †Département de Pédiatrie;
‡Département de Nutrition, Université de Montréal, Québec, Canada

The human scavenger receptor class B type I (SR-BI and splice variant SR-BII) plays a central role in HDL cholesterol metabolism and represents a candidate gene for a number of related diseases. We examined the genetic diversity of its coding and flanking regions in a sample of 178 chromosomes from individuals of European, African, East Asian (including Southeast Asian), Middle-Eastern as well as Amerindian descent. Nine of the 14 polymorphisms observed are new. Four of the five variants causing amino acid replacements, G2S, S229G, R484W, and G499R, are likely to affect protein structure and function. SR-BI/BII diversity is partitioned among 19 haplotypes; all but one interconnected by single mutation or a recombination event. Such tight haplotype network and the unusual geographic partitioning of this diversity, high not only in Africa but in East Asia as well, suggests its recent origin and possible effect of selection. Coalescent analysis infers a relatively short time to the most recent common ancestor and points to population expansion in Africa and East Asia. These two continents differ significantly in pairwise FST values, differing as well from a single cluster formed by Europe, Middle East and America. In the context of findings for similarly analyzed other loci, we propose that a selective sweep at the origin of modern human populations could explain the low level of ancestral SR-BI/II diversity. The unusually deep split between Africa and Asia, well beyond the Upper Paleolithic when inferred under neutrality, is consistent with subsequent geographical and demographic expansion favoring the accumulation of new variants, especially in groups characterized by large effective population sizes, such as Asians and Africans. The relevance of such partitioning of SR-BI/II diversity remains to be investigated in genetic epidemiological studies which can be guided by the present findings.

Introduction

SR-BI is a plasma membrane glycoprotein identified as the first cell surface high-density-lipoprotein (HDL) receptor (Acton et al. 1994; 1996), which plays a central role in the clearance of HDL cholesterol and in the regulation of its plasma levels (Ji et al. 1997; Rigotti et al. 1997; Jian et al. 1998; Varban et al. 1998). It mediates the binding of HDL, the selective uptake of cholesteryl esters without degradation of HDL protein moieties, and the bi-directional transfer of free cholesterol between HDL and cells. SR-BI recognizes a broad spectrum of ligands, including native low-density lipoprotein (LDL), modified lipoproteins, maleylated albumin, and anionic phospholipids (Acton et al. 1994; Calvo et al. 1997; Rigotti et al. 1997). It is involved in the suppression of atherosclerosis development (Braun et al. 1994; Calvo et al. 1997; Rigotti et al. 1997). It is involved in the regulation of hormone synthesis in streiodogenic tissues (Azhar and Reaven 2002), and in intestinal cholesterol absorption (Werder et al. 2001). Cholesterol metabolism during pregnancy and lactation (Smith et al. 1998), lipid delivery to the embryo (Hatzopoulos et al. 1998), and the recognition of senescent or apoptotic cells (Fukasawa et al. 1996; Muroa et al. 1997; Shiratsuchi et al. 1999) were also found to be influenced by this locus. SR-BI (-/-) mice displayed marked hypercholesterolemia characterized by impaired HDL clearance and dramatic sterol content decrease in adrenal tissue (Rigotti et al. 1997). In contrast, hepatic overexpression of SR-BI resulted in diminished HDL cholesterol content and enhanced delivery of HDL-associated lipid to hepatocytes and bile (Kozarsky et al. 1997).

The SR-BI gene is composed of 13 exons that are spread over 86 kb on chromosome 12q24.31–32; its splice variant SR-BII, obtained by skipping the 129-nucleotide-long exon 12, is expressed in rodents and humans (Webb et al. 1997; 1998). Certain hormones have been shown to modulate its splicing (Graf, Roswell, and Smart 2001), consistent with the notion that these isoforms may represent part of the HDL-cholesterol metabolism regulation mechanism. In the context of current efforts to identify and characterize polymorphisms conferring susceptibility alleles associated with disease risk or therapeutic response, genes implicated in cholesterol and lipid metabolism have drawn considerable attention (Clark et al. 1998; Fullerton et al. 2000; 2002). However, except for a few variants identified in individuals of European descent (Acton et al. 1999), there has been no systematic study of the SR-BI/BII locus. We present data on allelic diversity in its coding region and the resulting haplotypes. Our analysis suggests the possible involvement of selection in the evolutionary history of this locus, a finding that could be of potential relevance for genetic epidemiology.

Materials and Methods

DNA Samples

DNA samples were acquired from the Coriell Institute for Medical Research (Camden, N.J.) or obtained, on a non-nominative basis, from consenting adults who provided information about their ethnic, linguistic, and geographic origins (following a protocol approved by our Institutional Advisory Board). Individuals of European descent were represented by French-Canadians from Quebec (n = 9) and other European-Canadians (n = 9);
Table 1
Characteristics of PCR Amplification of SR-BI/II Segments

<table>
<thead>
<tr>
<th>Exon</th>
<th>Forward (5′→3′)</th>
<th>Reverse (5′→3′)</th>
<th>Annealing (°C)</th>
<th>Product Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−35: CATAAAAACCTGCGGCACTT</td>
<td>264: TAAAGCGCTGCTGGTAGT</td>
<td>63</td>
<td>299</td>
</tr>
<tr>
<td>2</td>
<td>−27: CAAGGTCTTGCCTGATCCT</td>
<td>+25: CACCCCTACCTGACAGCAGC</td>
<td>61</td>
<td>210</td>
</tr>
<tr>
<td>3</td>
<td>−26: GTGACCCACTGCTGCCCT</td>
<td>+54: AGACAGGGGAGGAGGCTG</td>
<td>64</td>
<td>222</td>
</tr>
<tr>
<td>4</td>
<td>−23: ACTAGTTGGCTTCTTGCCTT</td>
<td>+42: ACAGAGGCTGGAATCTTGG</td>
<td>61</td>
<td>269</td>
</tr>
<tr>
<td>5</td>
<td>−39: GCCTAAGGGCTTATGCTGGT</td>
<td>+90: CTTCTGGTTCCTGGCAGC</td>
<td>63</td>
<td>225</td>
</tr>
<tr>
<td>6</td>
<td>−30: GAGGGGCTTATGCTGGATC</td>
<td>+56: GAATTTGGGCGTATGACCT</td>
<td>62</td>
<td>202</td>
</tr>
<tr>
<td>7</td>
<td>−27: CCTCCTACATCTGTCTGGTC</td>
<td>+73: CTGGCAGAATATACCCTTCT</td>
<td>62</td>
<td>267</td>
</tr>
<tr>
<td>8</td>
<td>−54: GCACCTGCAATATACGTGCT</td>
<td>+51: GTGCTGCAACGGACAATCAC</td>
<td>62</td>
<td>224</td>
</tr>
<tr>
<td>9</td>
<td>−32: GGCGATCTCTCCACTGGTAGT</td>
<td>+64: CTCGACCACAGCAGCCATCA</td>
<td>62</td>
<td>170</td>
</tr>
<tr>
<td>10</td>
<td>−33: TGTGCTACAGCTCTGCTTGC</td>
<td>+57: AATGCCCAATGCGCTTCTA</td>
<td>60</td>
<td>142</td>
</tr>
<tr>
<td>11</td>
<td>−34: CCCAGCTTCGATTGATCTCA</td>
<td>+83: CAGAGGGTCGACGTCTGAG</td>
<td>62</td>
<td>264</td>
</tr>
<tr>
<td>12</td>
<td>−45: AGTGGCTTCAAGCTCAACT</td>
<td>+67: TAAAGCGCAGCATTCTACT</td>
<td>60</td>
<td>241</td>
</tr>
<tr>
<td>13A</td>
<td>−54: GTGGTTTGGTGACAGG</td>
<td>238: TACAAGCTCCCTTCACAGC</td>
<td>63</td>
<td>292</td>
</tr>
<tr>
<td>13B</td>
<td>195: TGTCGACAGCTCTGAGGAG</td>
<td>365: ACAGGCGGAAAGGCGGAC</td>
<td>63</td>
<td>171</td>
</tr>
<tr>
<td>13C</td>
<td>317: GTCCTGCTTCCTCCCTCC</td>
<td>525: CAGGTAGACCAAGGCAAAG</td>
<td>60</td>
<td>209</td>
</tr>
<tr>
<td>13D</td>
<td>456: TCCCTGTCGCAAATGCTGG</td>
<td>814: TTCAGCTGGGCGGCAATCAA</td>
<td>60</td>
<td>359</td>
</tr>
<tr>
<td>13E</td>
<td>677: CCTTCTCCGACCAGCTA</td>
<td>+33: TCAGCAAACAGCCTGATCA</td>
<td>60</td>
<td>314</td>
</tr>
</tbody>
</table>

a The number preceding the primer sequence describes the position of its 5′ nucleotide, upstream or downstream of the exon, into the introns indicated by minus or plus sign, respectively, or if exonic, the consecutive position in the SR-BI mRNA.

b 7.5% DMSO.

c 2.5% formamide.

Polymerase Chain Reaction Amplification

Polymerase chain reaction (PCR) primer pairs (table 1) were designed for exons 2–12 outside the exonic borders. For exon 1, where no downstream sequence (intron 1) was available, the reverse primer intrudes 22 nucleotides into the exon (GenBank entries Z22555, gi: 397606 and AC020773 for cDNA and genomic sequence, respectively). According to Cao et al. (1997), the nucleotide 143 positions upstream of initiation codon ATG was considered as a transcription start site, and hence the beginning of exon 1. For exon 13 (932 bp), five overlapping fragments (A to E) were used to cover the sequence with shorter DNA fragments better suited for Single-Strand Conformation Polymorphism (SSCP) analysis (Orita et al. 1989).

DNA segments were amplified by PCR in a 10-μl reaction using 5 ng of genomic sample, 0.75 U of Platinum Taq DNA polymerase (Invitrogen Canada, Inc.) in the Platinum PCR buffer containing 1.5 mM MgCl2 and 0.2 mM of each dNTP, and 2 μM of each primer in the presence of formamide or DMSO if so indicated in table 1. The PCR was performed for 35 cycles, each of 30 s at 94°C, 30 s at the annealing temperature (table 1), and 30 s at 72°C. For SSCP analysis, the PCR contained only 0.1 mM cold dCTP supplemented by 0.1 mM α32P-dCTP, 3,000 Ci/mmol (Amersham Biosciences).

SSCP Analysis, Sequencing, and Genotyping

Conformational DNA variants were analyzed by electrophoresis in a 6% polyacrylamide gel containing 10% glycerol (acylamide-bisacrylamide ratio of 29:1 and for exon 7 also 50:1) in TBE (0.089 M Tris-borate pH 8, 0.01 M EDTA). The PCR products were mixed (1:2) with 95% formamide, heated at 94°C for 5 min, and cooled on ice. Electrophoresis was carried out at room temperature for 16–21 h, at 400 V, in an S2 apparatus (Life Technologies). The dried gel was exposed using X-OMAT AR Kodak film. Gel mobility variants were analyzed by dyeoxy-sequencing (Thermo Sequenase Radiolabeled Terminator Cycle Sequencing Kit, USB Corporation, Cleveland, Ohio) according to the manufacturer’s protocol. Partial sequences in a subset of the analyzed fragments were determined in the common chimpanzee (DNA sample from BIOS Laboratories) in order to assist the assignment of the ancestral allele in polymorphisms 1–4, 9, 12, and 13. Because of characteristic gel mobility profiles, alleles were unambiguously assigned for both homozygotes and heterozygotes individuals. The genotypes obtained in this way were subsequently used to solve the underlying haplotypes.

Haplotype Solving and Networks

Of the 89 individuals examined, 60 were either homozygous or heterozygous at a single site, yielding unambiguous haplotype solutions. Furthermore, 27 individuals were heterozygous at more than one site and 2 had missing genotypes at position 4 or positions 1–3, respectively. All genotypes were analyzed using the software PHASE, version 1.0 (Stephens, Smith, and Donnelly 2001). Repeated analyses led to stable solutions except for four
genotypes consisting of combinations of one known and a new or inferred haplotype. Preference was given to combinations involving known haplotypes of higher regional population frequency. The alternative solutions are reported later (table 3) to indicate that these minor haplotypes are not inferred without ambiguity inherent in the data. A maximum parsimony network of haplotypes was constructed manually, first from the haplotypes connectable by mutations only, and then by addition of the recombinant haplotypes (recurrent mutations were considered unlikely).

### Table 2

**Diversity Statistics in SR-BI/II**

<table>
<thead>
<tr>
<th></th>
<th>Europe</th>
<th>Middle East</th>
<th>Africa</th>
<th>East Asia</th>
<th>America</th>
<th>World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of chromosomes, n</td>
<td>36</td>
<td>46</td>
<td>36</td>
<td>18</td>
<td>42</td>
<td>178</td>
</tr>
<tr>
<td>Number of polymorphic sites, S</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Number of haplotypes, k</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

### Population statistics

- Haplotype diversity, G (SD) 0.62 (0.047)
- Nuclear diversity (%), π = Π/L (SD) 0.67 (0.05)
- θ (based on Πi), θRi = Πi (SD) 0.84 (0.041)
- θ (based on S), θRi = S/S2Ri (i/i) (SD) 0.89 (0.044)
- Tajima’s estimator D 0.0109
- Fu’s estimator Fs 1.28

### Statistical Analyses

The software Arlequin v.2.0 (Schneider, Roessli, and Excoffier 2000) was used to obtain summary statistics of genetic diversity (see Results) and to test for selective neutrality (Tajima 1989; Fu 1997). Genetic differentiation among population samples was quantified by the test statistic FST (Wright 1951), computed according to Weir and Cockerham (1984), as implemented in Arlequin, per site or locus, for pairs of continental populations or for the total sample subdivided into five continental groups.

### Table 3

**SR-BI/II Haplotypes and Their Geographic Distribution**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Europe</th>
<th>Middle East</th>
<th>Africa</th>
<th>East Asia</th>
<th>Americas</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>G C C G T</td>
<td>17</td>
<td>21</td>
<td>6</td>
<td>3</td>
<td>19</td>
<td>66</td>
</tr>
<tr>
<td>H2</td>
<td>. . . .</td>
<td>15</td>
<td>16</td>
<td>12</td>
<td>4</td>
<td>14</td>
<td>61</td>
</tr>
<tr>
<td>H3</td>
<td>A . . . .</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>. . . .</td>
<td>0.201</td>
<td>0.064</td>
<td>0.012</td>
<td>0.005</td>
<td>0.123</td>
<td>0.000</td>
</tr>
<tr>
<td>H5</td>
<td>. . . .</td>
<td>3.44 (1.41)</td>
<td>4.15 (1.65)</td>
<td>5.28 (1.75)</td>
<td>4.94 (1.65)</td>
<td>4.60 (1.71)</td>
<td>6.50 (2.20)</td>
</tr>
<tr>
<td>H6</td>
<td>. . . .</td>
<td>0.77</td>
<td>0.73</td>
<td>2.52</td>
<td>2.19</td>
<td>1.32</td>
<td>3.06</td>
</tr>
<tr>
<td>H7</td>
<td>. . . .</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>H8</td>
<td>(H1/H3)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H9</td>
<td>(H7/H10)</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>H10</td>
<td>. . . .</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H11</td>
<td>(H1/H6)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>H12</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>H13</td>
<td>(H2/H4)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H14</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H15</td>
<td>(H1/H14)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H16</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H17</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H18</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>H19</td>
<td>. . . .</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

| Total | 36 | 46 | 36 | 18 | 42 | 178 |

- Recombinant haplotypes were given alias names in parentheses, indicating putative parental haplotypes.
- Dot denotes agreement with H1 nucleotide at respective site.
- Observed only among alternative haplotype solutions and not retained for further analysis (see Materials and Methods).
The genomic clone NCBI gi: 22003948, AC126309, and its adjacent and overlapping contig gi: 22058877, NT_035239.1, respectively.

of the total gene length of about 86 kb was obtained by aligning the SR-BI mRNA’s (GeneBank gi: 397606) upstream and downstream portions with the transcription start site within the mRNA sequence or, if in the intronic sequence (int), upstream/downstream of the corresponding exon. An estimate which leads to a stop translation in exon 13 (Webb et al. 1998). Positions of the polymorphic sites numbered 1 to 14 are either given as a distance from the amplicons. Fragments 13A to 13E represent five amplicons covering exon 13. SR-BII differs from SR-BI by alternative splicing, omitting exon 12.

Pairwise $F_{ST}$’s between the five groups were represented by multidimensional scaling using STATISTICA v. 6.0.

Coalescence analysis was carried out according to the method of Griffiths and Tavare (1994) using genetree (v. 8.3) on the full data set (excluding recombinant haplotypes—i.e., 12 out of 178 chromosomes in the analyzed sample); maximum likelihood estimates of $\theta$ (i.e., $\theta_{ML}$), the time to the most recent common ancestor (TMRCA), and the ages of mutations were obtained conditional on the haplotype tree, assuming an infinite-sites model, constant population size, and random mating. The number of iterations per run was sufficiently large for estimates to remain constant over repeated runs, which only differed in the random number seed. Estimates of $\theta_{ML}$ were also obtained under a model of exponential growth, concurrently yielding an estimate of growth rate $\beta$. Particularly in this model the population size exponentially declines backward in time at rate $\beta$ from a current size $N_e(0)$, such that the size of the population at time $t$ is $N(t) = N_e(0) e^{-\beta t}$ (note that $t = g/2N_e(0)$, where $g$ is the number of generations ago). To calculate effective population sizes, we used the mutation rate estimate of Fan et al. (2002) of $1.04 \times 10^{-8}$ per nucleotide per generation (or $5.2 \times 10^{-10}$ per nucleotide per year), obtained from a comparison of 108-kb human expressed sequence tags (ESTs) with the corresponding great and lesser ape sequences. This rate compares well with the mutation rate of $10^{-9}$ per nucleotide per year based on the divergence of 0.15 between human and bovine SR-BI/II cDNAs, taking into account the slowing down of the mutation rate and extended generation time in primate lineages (Koop et al. 1989; Yi, Ellsworth, and Li 2002). Note that use of the value of $10^{-9}$, twice as large as the Fan et al. (2002) estimate, would result in a twofold decrease of the reported $N_e$ and, consequently, the TMRCA and mutation age estimates. From the genetic map of Kong et al. (2002), we obtained an average recombination rate of about 3.1% per Mb per generation.

**Interspecies Comparative Analysis**

Amino acid sequence of the human protein was compared to the corresponding sequences of rat, mouse, bovine, Chinese hamster, and pig (GenBank entries NP 113729, NP 058021, AAB70920.1, AAA61572, and AAL75567.1, respectively) to assess evolutionary conservation of amino acid substitution polymorphisms.

**Results**

**SR-BI/II Variants**

We analyzed the genetic diversity of the quasi-totality of the SR-BI/II coding region (1,624 bp), both 5’ and 3’ untranslated regions (980 bp), as well as 627 bp of the flanking introns (fig. 1). Twelve transitions and two transversions were observed in a worldwide sample of 178 chromosomes. Eleven of these sites were in the coding regions, five of which caused amino acid replacements, one was in 3’-UTR, and two were intronic (table 4). Three polymorphisms were seen in at least four continents, four occurred in two continents, and seven occurred in a single continent only, suggesting their recent origin. Tajima’s test statistic $D$ (Tajima 1989) with $P = 0.071$ for the total (world) sample (table 2), bearing in mind its conservative assumption of no recombination, might be considered suggestive of non-neutrality or of population expansion. This would be consistent with the highly significant $F_{ST}$ statistics (table 2) in the world, in Africa, and in East Asia (Fu 1997; Ramos-Onsins and Rozas 2002).

Of the 19 haplotypes deduced, H1 to H19, 14 are due to mutations, whereas 5 haplotypes resulted from single crossovers (table 3). The number of haplotypes (14) observed that are due to mutation is unusually large—i.e., almost equal to the maximum number of 15 haplotypes $(S + 1)$ that can arise from 14 biallelic sites, not all of which are expected to be found in a finite population at the same time. Therefore, our observation may indicate the effect of selection, or a dramatic reduction of genetic drift, due

**Fig. 1.—Schematic representation of the SR-BI/II gene.** Horizontal bars correspond to 13 exons with extruding intronic fragments when present in the amplicons. Fragments 13A to 13E represent five amplicons covering exon 13. SR-BII differs from SR-BI by alternative splicing, omitting exon 12.
Table 4

Frequencies and Geographic Distribution of SR-BI/II SNPs

<table>
<thead>
<tr>
<th>Mutation/Variant</th>
<th>Exon/Intron</th>
<th>Nucleotide</th>
<th>Amino Acid</th>
<th>Change&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Europe (n = 36)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Middle East (n = 40)</th>
<th>Africa (n = 36)</th>
<th>East Asia (n = 18)</th>
<th>Americas (n = 42)</th>
<th>Total (n = 178)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
<td>1</td>
<td>G147A</td>
<td>G2S</td>
<td></td>
<td>.083 (.046)</td>
<td>.130 (.050)</td>
<td>.056 (.038)</td>
<td>.071 (.040)</td>
<td>.079 (.020)</td>
<td></td>
</tr>
<tr>
<td>2 1</td>
<td>1</td>
<td>C155T</td>
<td></td>
<td></td>
<td>.022 (.022)</td>
<td>.167 (.062)</td>
<td>.028 (.027)</td>
<td></td>
<td>.039 (.015)</td>
<td></td>
</tr>
<tr>
<td>3 1</td>
<td>1</td>
<td>C206T</td>
<td></td>
<td></td>
<td>.028 (.027)</td>
<td>.139 (.058)</td>
<td>.022 (.022)</td>
<td></td>
<td>.006 (.006)</td>
<td></td>
</tr>
<tr>
<td>4 3</td>
<td>5</td>
<td>G546A</td>
<td>V135I</td>
<td></td>
<td>.111 (.074)</td>
<td>.056 (.054)</td>
<td>.111 (.074)</td>
<td></td>
<td>.006 (.006)</td>
<td></td>
</tr>
<tr>
<td>5 4</td>
<td>6</td>
<td>C644T</td>
<td></td>
<td></td>
<td>.028 (.027)</td>
<td>.022 (.022)</td>
<td>.022 (.022)</td>
<td>.389 (.115)</td>
<td>.095 (.045)</td>
<td>.062 (.018)</td>
</tr>
<tr>
<td>6 7</td>
<td>5</td>
<td>T+16A</td>
<td></td>
<td></td>
<td>.111 (.074)</td>
<td>.194 (.066)</td>
<td>.278 (.106)</td>
<td></td>
<td>.416 (.037)</td>
<td></td>
</tr>
<tr>
<td>7 5</td>
<td>6</td>
<td>A828G</td>
<td>S229G</td>
<td></td>
<td>.111 (.074)</td>
<td>.011 (.008)</td>
<td></td>
<td></td>
<td>.011 (.008)</td>
<td></td>
</tr>
<tr>
<td>8 7</td>
<td>8</td>
<td>C1046T</td>
<td></td>
<td></td>
<td>.028 (.027)</td>
<td>.194 (.066)</td>
<td>.278 (.106)</td>
<td>.476 (.077)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 8</td>
<td>10</td>
<td>C1193T</td>
<td>G644</td>
<td></td>
<td>.500 (.083)</td>
<td>.222 (.074)</td>
<td>.139 (.058)</td>
<td></td>
<td>.304 (.013)</td>
<td></td>
</tr>
<tr>
<td>10 10</td>
<td>11</td>
<td>C1722T</td>
<td>R484W&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td>.022 (.022)</td>
<td>.022 (.022)</td>
<td>.167 (.062)</td>
<td></td>
<td>.034 (.013)</td>
<td></td>
</tr>
<tr>
<td>12 12</td>
<td>13</td>
<td>G1745A</td>
<td></td>
<td></td>
<td>.333 (.111)</td>
<td>.095 (.045)</td>
<td></td>
<td></td>
<td>.056 (.017)</td>
<td></td>
</tr>
<tr>
<td>13 13</td>
<td>14</td>
<td>G1767A</td>
<td>G499R&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td>.022 (.022)</td>
<td>.022 (.022)</td>
<td></td>
<td></td>
<td>.006 (.006)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Position of nucleotide variants in exons is given according to SR-BI cDNA sequence numbering, starting 143 nucleotides upstream of the ATG initiation codon.

<sup>b</sup> n denotes the number of chromosomes analyzed.

<sup>c</sup> Amino acid position in SR-BII.

<sup>d</sup> SD, standard deviation.

either to a very large population size or to its rapid growth (see Beheregrady et al. 2003). Finding all intermediate haplotypes in a relatively limited population sample argues against a very large population size, thus suggesting rapid population growth and/or selection. The 19 haplotypes were easily connected within a minimum-step spanning network (fig. 2). Two major haplotypes, H1 and H2, which differ by one mutation (C1193T), account for 71% of the global sample. The direction of the 1193 mutation is from C to T, i.e., from H2 to H1, because C, shared with the chimpanzee, can be considered ancestral. Although H1 and H2 occur at the same worldwide frequency, only two mutations, 6 and 11, appear to have occurred on the background of H1, whereas 10 arose on the background of H2. More haplotypes derive from H2 than from H1, including haplotype branch 13–3–4 that occurs locally in Sub-Saharan Africa and haplotype branch 5–12–7 from East Asia. Continentally shared haplotypes can be ascribed to gene flow either from East Asia to the Americas or from Africa through the Middle East to Europe (fig. 2).

F<sub>ST</sub> and Population Differentiation

An overall F<sub>ST</sub> estimate of 9.2% is in the lower range of values (10% to 15%) typically observed between continental groups with protein or autosomal DNA markers (e.g., Cavalli-Sforza, Menozzi, and Piazza 1994; Barbujani et al. 1997; Jorde et al. 2000; Fan et al. 2002). The contribution of individual polymorphic sites was uneven, with only half of the sites showing F<sub>ST</sub> values significantly different from zero. In the overall apportionment of genetic diversity across continents, measured through pairwise F<sub>ST</sub>’s and represented by multidimensional scaling (fig. 3), Africa and East Asia appear to differ significantly from Europe, the Middle East, and the Americas, consistent with local networks of haplotypes shown in figure 2.

Amino Acid Polymorphisms

Sites polymorphic at the human protein level (table 4) were compared to their orthologous positions in three rodent and two artiodactyl species (see Materials and Methods). G2, S229, and G499 were found in all aligned sequences, indicating evolutionary conservation across three (i.e., including Primates) mammalian orders. In contrast, V135 was found to be replaced by I in the Chinese hamster. At position 484, however, with R and derived W in the human protein, Q is found in the orthologous site in all nonprimate species compared. Indeed, the V to I substitution is known to be “conservative,” i.e., it rarely affects the protein function due to structural similarity of these amino acids. Throughout evolution it occurs with a relative frequency of 33, as measured by the PAM (point accepted mutation) matrix. This is in contrast to the relative frequency of 16 and 21 for G to S and S to G replacements, respectively, of 10 for Q to R, 2 for R to W, and zero for both G to R and Q to W substitutions (Felsenstein 2004). The latter two thus represent “nonconservative” amino acid substitutions that are likely to affect the SR-BI/II structure and function. We will see that this might also be the case of less dramatic amino acid change at sites 2 and 229.

Age of Mutations

We used a coalescent model to infer the time scale of the origin and evolution of the variation within SR-BI/II locus. The underlying gene tree used in this analysis (fig. 4) was rooted by assuming that haplotype H2 was ancestral (table 3); five recombinant haplotypes representing 12 of the 178 chromosomes analyzed were not included in this analysis (see Materials and Methods). Given \( \theta = 4N_e \mu \), and \( \mu = 3.5 \times 10^{-5} \) as mutation rate per analyzed SR-BI/II segment per generation, we evaluated...
Ne at 8,500, 17,500, and 21,900 from $h/C_5$, $h_S$ and $h_{ML}$, respectively, comparable to values reported in the literature (Li and Sadler 1991; Harding et al. 1997; Jaruzelska et al. 1999; Fan et al. 2002). For a generation time of 20 years, assuming $h$ to be known and setting it equal to $h_{ML}$, we obtained an estimate of TMRCA of 1.04 (standard deviation of 0.36) in coalescence units (or 2 $N_e$ generations; fig. 4), corresponding to a TMRCA of 910 ± 320 Kyr (thousand of years). Using $h/C_5$ yields a TMRCA of 1.55 ± 0.5 coalescence units, which corresponds to a time depth of 530 ± 170 Kyr, and thus a shallower tree than the one based on $h_{ML}$. Time estimates based on $h_S$ are between those using $h_{ML}$ and $h_{II}$ (see $\theta$ values in table 2). From this point forward, we will only refer to the values obtained based on $\theta_{ML}$ (table 2). The age of mutation 9 is 0.42 ± 0.25 in 2$N_e$ generations (fig. 4) or 370 ± 220 Kyr. The estimates for mutations 13, 5, 1, and 12 were 250 ± 210, 250 ± 200, 140 ± 135, and 90 ± 75 Kyr, respectively. Other mutations appeared younger, most less than 50 or even 10 Kyr old (fig. 4, filled circles). On a relative scale, the oldest mutation, 9, which splits the diversity between H2-derived and H1-derived branches, occurs closer to the present than the half-time to the TMRCA. In contrast to a multiplicity of H2-derived branches, only two branches, those marked by mutations 6 and 11, follow mutation 9. This suggests that mutation 9 could be younger than it appears, either because of recent rapid growth or because of a founder effect in its carrier population(s) of Europe.

![Fig. 2.—SR-BI/II haplotype networks for the total sample and by groups of continental populations. Each haplotype is represented by a circle, whose square radius (surface) is proportional to its population frequency. H2 is the likely ancestral haplotype. Solid lines represent mutation events creating polymorphisms 1 to 14 (see table 4), and broken lines indicate recombinations leading to new haplotypes.](image-url)
the Middle East, and the Americas (note that the estimated mutation age is a function of the new allele’s frequency). In Africa and East Asia the prevalent H2 gave rise to much greater haplotype diversity. Are these latter continental populations relatively older or were the effective sizes of their human populations relatively larger?

In fact, genetree provides a better representation of data under a growth scenario (see table 2 for joint estimates of $\beta$ and $\theta_{ML}$) than under a constant population size model. This is seen in the world sample, as well as in Africa and East Asia, where a star-like phylogeny is suggested by the shallow shape of the tree, with all but one of the intermediary haplotypes present (fig. 2), and most of the mutations (and recombinations) positioned at the tip of the branches (Donnelly 1996; Harpending et al. 1998). Although the heights of the resulting trees (in $2N_e(0)$ units) are smaller for a growth model (fig. 4), the difference between TMRCAs in years is less dramatic because the jointly estimated $\theta_{ML}$s are also larger, as are the resulting $N_e$’s. Essentially identical TMRCAs, of 680 Kyr, 684 Kyr, and 734 Kyr were evaluated for the world, Africa, and East Asia, respectively. The ages of mutation 9 (415 Kyr) and all other polymorphisms, when evaluated in the world sample, were slightly higher than under a constant size model (fig. 4, open symbols).

**Discussion**

In this study, 9 of 14 single-nucleotide polymorphisms (SNPs) found are previously unreported; five variants in the coding region caused amino acid replacements. The nature of these substitutions and their interspecies comparison suggest that some of these polymorphisms, and particularly at amino acid positions 484 and 499, might indeed modify protein structure, and thus affect its function. In addition, a correlation was reported between the serine allele (variant G2S), an increased level of HDL and a decreased concentration of LDL cholesterol (Acton et al. 1999). We found this G2S polymorphism in all populations except for East Asia, consistent with the findings reported by Hong et al. (2002). Therefore, a possible role of this polymorphism in cholesterol-related disorders needs to be investigated further in appropriate cohorts. In contrast, V135I which is not expected to have any significant consequence on protein structure showed no association with plasma lipid and circulating lipoprotein levels, or with the body mass index (Acton et al. 1999). In this respect, G to S and S to G replacements, although less frequent than V to I replacements, represent relatively “permissive” substitutions. Still, both G2 and S229 were found to be conserved in three mammalian orders examined, which may explain the possible functional importance and the effect of G2S polymorphism above as observed by Acton et al. (1999). The importance of S229 may be ascribed to its sequence context. Because the amino acid sequence at positions 227–230 (NISR) defines the consensus sequence for N-glycosylation, G229 renders this putative glycosylation site nonfunctional. The absence of glycosylation in the predicted extracellular domain of the SR-BI/II protein might affect the selective uptake of HDL cholesteryl ester as well as the bi-directional free cholesterol flux. This relationship is to be examined in the East Asian population, where the G229 allele is the most prevalent.

The silent polymorphism C1193T in exon 8, which is the oldest SNP and effectively partitions the SR-BI/II variability almost in half (table 4), is associated with low plasma LDL cholesterol in healthy Caucasian women (Acton et al. 1999) as well as with high plasma HDL.
cholesterol and a lower risk of coronary artery disease in Koreans (Hong et al. 2002). As a result, all functional variability that arose on the background of H1 and H2 will be in linkage disequilibrium with the 1193T allele and the C allele, respectively. Hence, these associations do not necessarily indicate functional significance of this apparently silent T to C polymorphism, but rather point to its linkage with other polymorphisms such as the observed amino acid substitutions that are more likely to affect the protein function. Knowledge of the haplotype network can therefore assist in the design of an epidemiological survey (i.e., how to choose informative markers for genotyping) and in the subsequent functional interpretation of the underlying variation.

The nucleotide diversity of SR-BI/II is similar to that of other genomic segments, particularly coding or expressed regions (Cargill et al. 1999; Fan et al. 2002; Nakajima et al. 2002; Schneider et al. 2003). We observed two common frequent haplotypes (H1 and H2) that dominate over a flat distribution of rare haplotypes due to novel mutations or recent recombinations. However, the total number of 19 haplotypes appears low for a gene spread over 86 kb on the physical map and 0.26 cM on the genetic map (Kong et al. 2002). With a recombination rate of 0.0026 (1/385) per gene segment per generation, a significant fraction of which is informative (estimated at 17% in the world sample—data not shown; see appendix in Zietkiewicz et al. 2003), a greater number of recombinant SR-BI/II haplotypes would be expected given the depth of the coalescence tree. In contrast, the observed recombinants can be considered relatively young given their frequency (fig. 2).

Another particularity is the geographic distribution of the genetic diversity of the SR-BI/II locus. Nucleotide diversity among Asians is equal to or even exceeds that in Africa (0.059% and 0.044%, respectively), an observation rarely reported in other loci (Harding et al. 1997; Jaruzelska et al. 1999). At the haplotype level, both Africa and East Asia are significantly more variable than the other continental samples (table 2). All this could suggest that African and Asian lineages were similarly ancestral to present-day populations. The estimated ages of mutations 5 and 13, East Asian and African specific, respectively, exceed 200 Kyr, and, if taken literally, would indicate an unusually ancient split between the East Asian and African lineages (fig. 4). Continentially restricted recombinant haplotypes H15 and H9 further emphasize isolation of these continental groups. Genetic separation between Africa and East Asia is also reflected in the $F_{ST}$ analysis (fig. 3). Even if the genetic history of SR-BI/II followed a trajectory different from that predicted by the recent out-of-Africa model (e.g., Harding et al. 1997; Hammer et al. 1998; Cruciani et al. 2002), one would still expect to observe a recent African influence in East Asia from the Upper Palaeolithic to the present day (Satta and Takahata 2002). In contrast, the deep split between East Asia and Africa seen in our data may actually be as recent as the late Middle or Upper Palaeolithic, and for several reasons.

First, because the haplotypes observed here represent a tight network with no intermediate missing forms (fig. 2).

This could indicate a small role played by genetic drift in the evolution of SR-BI/II genetic diversity as observed today (see e.g., Behegercaray et al. 2003). Indeed, over time, and in a finite population, new haplotypes are created while others become extinct. Thus, at a given historical moment, those haplotypes that have arisen through multiple mutation and recombination events represent only a subset of all allelic combinations as is seen in the PDHA1 locus (Harris and Hey 1999), in beta-globin (Harding et al. 1997), dystrophin (Labuda, Zietkiewicz, and Yotova 2000), CYP1A2 (Wooding et al. 2002), subterminal 16p (Alonso and Armour 2001), or APOE (Fullerton et al. 2000). Although tight networks were observed (1) in very short genomic segments (Jaruzelska et al. 1999; Jin et al. 1999), (2) in segments of low recombination rate (Kaessmann et al. 1999), or (3) in segments under selection (Smirnova et al. 2001), in SR-BI/II, extending over 86 kb and with the recombination rate exceeding threefold the genomic average, explanations (1) and (2) hardly apply. Second, when evaluated through a simple model of constant population size, the overall time depth of the history of SR-BI/II (2.08 $Ne$ generations—fig. 4) is about half that expected (4 $N_e$) for a neutrally evolving autosomal locus. The SR-BI/II tree appears even shallower than those of X-chromosome loci (Harris and Hey 1999; Jaruzelska et al. 1999; Verrelli et al. 2002). Third, population expansion provides a better description of the data than a constant population size model, as seen under the coalescence analysis and based on statistics such as Fu’s $F_{S}$ and $Tajima’s D$ (table 2). However, it is noteworthy that slowly accumulating nucleotide variation in nuclear loci has only rarely provided evidence of expansion (Thomson et al. 2000; Alonso and Armour 2001), as the expansion period (Klein 1999) was relatively recent and presumably too short on an evolutionary scale for new variants to accumulate. Thus, the evidence of population expansion in coalescence analysis may likely be due, in fact, to selection as supported by Tajima’s $D$ and Fu’s $F_{S}$. Fourth, there is a striking absence of significant nucleotide variability in SR-BI/II that is ancestral to the divergence of continental populations, as well as a shortage of haplotype diversity, due to recombination, given the spread of genetic distances within the locus. Selection is therefore plausible, and if this is the case, the time estimates based on the neutral model are certainly misleading. Considering all findings and in comparison with other similarly analyzed loci, which were “shaped” by the same demogenetic processes, it is conceivable that selection affected the genetic history of the SR-BI/II locus in the lineage leading to modern humans. In summary, we propose that rather than supporting an ancient split between Africa and Asia, our data can be interpreted assuming that the SR-BI/II variability is relatively young and accrued during the period following the recent out-of-Africa expansion (Cavalli-Sforza, Menozzi, and Piazza 1994). A selective sweep at the origin of modern humans could have been responsible for the disappearance of its ancestral diversity. Because the local recombination rate is relatively high, it is possible that SR-BI/II itself could have been under the selection process rather than being affected by the genetic hitchhiking that derived from selection involving a linked gene.
Acknowledgments

We are grateful to all individuals who kindly consented to provide DNA for this study. We thank Patrick Beaulieu, Dominik Gehl, and Vanja Yotova for their help; Alan Lovell for the English corrections; Dominika Kozubska for her secretarial assistance; and our reviewers for their very helpful comments. D.S. is a scholar of the Fonds de la Recherche en Santé du Québec. This work was supported by the Fondation de Hôpital Ste-Justine and in part by Valorisation Recherche Québec and the Canadian Institutes of Health Research grant (MOP-12782 to D.L.).

Literature Cited


Alonso, S., and J. A. Armour. 2001. A highly variable segment of their help; Alan Lovell for the English corrections; Patrick Beaulieu, Dominik Gehl, and Vania Yotova for


Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland.


David B. Goldstein, Associate Editor

Accepted December 5, 2003