Nucleotide Diversity and Haplotype Structure of the Human Angiotensinogen Gene in Two Populations

Toshiaki Nakajima,1 Lynn B. Jorde,2 Tomoaki Ishigami,2,3 Satoshi Umemura,3 Mitsuru Emi,4 Jean-Marc Lalouel,2 and Ituro Inoue1

1Division of Genetic Diagnosis, The Institute of Medical Science, The University of Tokyo, Tokyo; 2Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City; 3Internal Medicine, Yokohama City University, Yokohama; and 4Department of Molecular Biology, Institute of Gerontology, Nippon Medical School, Kawasaki, Japan

Variation in the angiotensinogen gene, AGT, has been associated with variation in plasma angiotensinogen levels. In addition, the T235M polymorphism in the AGT product is associated with an increased risk of essential hypertension in multiple populations, making AGT a good example of a quantitative-trait locus underlying susceptibility to a common disease. To better understand genetic variation in AGT, we sequenced a 14.4-kb genomic region spanning the entire AGT and identified 44 single-nucleotide polymorphisms (SNPs). Forty-two SNPs were observed both in 88 white and in 77 Japanese unselected subjects. Six major haplotypes accounted for most of the variation in this region, indicating less allelic complexity than in many other genomic regions. Although the two populations were found to share all of the major AGT haplotypes, there were substantial differences in haplotype frequencies. Pairwise linkage disequilibrium (LD), measured by the D', r², and d² statistics, demonstrated a general pattern of decline with increasing distance, but, as expected in a small genomic region, individual LD values were highly variable. LD between T235M and each of the other 39 SNPs was assessed in order to model the usefulness of LD to detect a disease-associated mutation. Among the Japanese subjects, 13 (33%) of the SNPs had r² values >0.1, whereas this statistic was substantially higher for the white subjects (occurring in 35/39 [90%]). LD between a hypertension-associated promoter mutation, A-6G, and 39 SNPs was also measured. Similar results were obtained, with 33% of the SNPs showing r² > 0.1 in the Japanese subjects and 92% of the SNPs showing r² > 0.1 in the white subjects. This difference, which occurs despite an overall similarity in LD patterns in the two populations, reflects a much higher frequency of the M235-associated haplotype in the white sample. These results have important implications for the usefulness of LD approaches in the mapping of genes underlying susceptibility to complex diseases.

Introduction

Much attention is now focused on the identification of susceptibility genes underlying complex diseases, through whole-genome linkage disequilibrium (LD) mapping with single-nucleotide polymorphisms (SNPs). The feasibility of such studies is currently under debate and depends explicitly on the persistence of LD between SNPs and causal mutations (Risch and Merikangas 1996; Collins et al. 1997; Kidd et al. 1998, 2000; Tishkoff et al. 1998, 2000b; Zavattari et al. 2000). Furthermore, the extent of LD varies considerably among different populations, reflecting the effects of population structure and history (Laan and Pa¨a¨bo 1997; Kidd et al. 1998, 2000; Tishkoff et al. 1998, 2000b; Zavattari et al. 2000). Finally, the presence of several disease-predisposing alleles within a susceptibility locus, each in association with a different background haplotype, can seriously compromise the ability of LD to locate the susceptibility locus (Xiong and Guo 1998). Considering the potential effects of these and other factors, it is not surprising that simulations and empirical studies have arrived at highly disparate results regarding both the expected extent of LD in the human genome and the resultant SNP density required for successful LD studies (Jorde 1995; Collins et al. 1999; Kruglyak 1999; Bonnen et al. 2000; Eaves et al. 2000; Moffatt et al. 2000; Abecasis et al. 2001; Reich et al. 2001; Stephens et al. 2001). Because of their important implications for the design of gene-mapping studies, these issues need to be resolved by additional empirical data.

AGT (MIM 106150) represents one of the few genes
et al. 2000; Pan et al. 2000; Rankinen et al. 2000; Rice et al. 2000; Sato et al. 2000), but not all (Brand et al. 1998; Bengtsson et al. 1999; Ni et al. 1999; Taittonen et al. 1999; Kato et al. 2000; Larson et al. 2000; Province et al. 2000). This inconsistency may reflect differences in phenotype definition, lack of statistical power, population history or structure, the effects of other loci, and the varying effects of several disease-predisposing variants within AGT (Corvol et al. 1999; Lalouel 2001). Nevertheless, several major meta-analyses have confirmed a significant association between AGT variation and hypertension, with a combined relative risk of ~1.2 for the T235 allele (Kunz et al. 1997; Kato et al. 1999; Staessen et al. 1999). AGT thus represents an important

### Table 1

<table>
<thead>
<tr>
<th>Polymorphism</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP: 1</td>
<td>5'-CAAGTATTTTTTAGGAGGAGTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 2</td>
<td>5'—CAAGTATTTTTAGGAGGAGTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 3</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 4</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 5</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 6</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 7</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 8</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 9</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 10</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 11</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 12</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 13</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 14</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 15</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 16</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 17</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 18</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 19</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 20</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 21</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 22</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 23</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 24</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 25</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 26</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 27</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 28</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 29</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 30</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 31</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 32</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 33</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 34</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 35</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 36</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 37</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 38</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 39</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 40</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 41</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
<tr>
<td>SNP: 42</td>
<td>5'-CTGCTCTTCACTGCTCTATC-3'</td>
</tr>
</tbody>
</table>

in which genetic variation has been shown to be associated with measurable variation in an endophenotype (plasma angiotensinogen) and in a biomedically relevant phenotype, hypertension (Juenenart et al. 1992). In previous studies, we have reported that two common polymorphisms, T235M and A-6G, are significantly associated with essential hypertension (EHT) (MIM 145500) (Inoue et al. 1997; Juenenart et al. 1997). The T235 allele is in nearly complete LD with A(-6) and is associated with higher plasma angiotensinogen levels than are the M 235 and G(-6) alleles. These results have been replicated in many other studies (Iso et al. 2000; Pan et al. 2000; Rankinen et al. 2000; Rice et al. 2000; Sato et al. 2000), but not all (Brand et al. 1998; Bengtsson et al. 1999; Ni et al. 1999; Taittonen et al. 1999; Kato et al. 2000; Larson et al. 2000; Province et al. 2000). This inconsistency may reflect differences in phenotype definition, lack of statistical power, population history or structure, the effects of other loci, and the varying effects of several disease-predisposing variants within AGT (Corvol et al. 1999; Lalouel 2001). Nevertheless, several major meta-analyses have confirmed a significant association between AGT variation and hypertension, with a combined relative risk of ~1.2 for the T235 allele (Kunz et al. 1997; Kato et al. 1999; Staessen et al. 1999). AGT thus represents an important
Subjects and Methods

Subjects

Seventy-seven Japanese individuals unselected for disease status were recruited from outpatient clinics in Yokohama City. Informed consent was obtained from each subject, and the study was performed with the approval of the Ethical Committee of Yokohama City University. Blood samples were collected for isolation of genomic DNA. The 88 white subjects are unrelated individuals from the Utah subset of the CEPH collection.

Isolation of PAC/Bacterial Artificial Chromosome (BAC) Clone and Genome Sequence of Human and Chimpanzee AGT

A bacteriophage PAC library containing human genomic DNA pooled in a three-dimensional structure (Genome Systems) was screened for the AGT clone. The PAC library was screened by the method described elsewhere (Nakajima et al. 2000), by use of two oppositely oriented oligonucleotides (5'-AGGCTGTACAGGGCTGTAGT-3' and 5'-GCCTTACCTTGGAAGTGGACGTA-3').

A high-density hybridization filter for chimpanzee genomic DNA is available from BAC/PAC Resources, Children's Hospital Oakland Research Institute. The filters were hybridized with digoxigenin-labeled (randomly primed; Roche) probes on exon 2 of AGT. Escherichia coli bearing the clones were cultured, and BAC/PAC DNA was isolated, as described elsewhere (Nakajima et al. 2000).

Promoter and exon sequences were obtained from GenBank (accession numbers NM_000029 and X15323). Intron sequences were determined from a PAC genome clone containing AGT, by direct primer walking across the gaps. Sequencing was performed by BigDye Terminator cycle sequencing using an ABI 377 Prism automated DNA sequencer (Applied Biosystems). Interspersed repeats in the gene were identified by RepeatM asker (see the RepeatM asker Documentation web site).

Identification of SNPs

Overlapping primer sets covering the genome sequence of AGT were designed on the basis of size and overlap of PCR amplicons (table 1). Genomic DNA was subjected to PCR amplification followed by sequencing using the BigDye Terminator cycle. Polymorphisms were identified by the comparison of sequences from 72 chromosomes (36 from Japanese and 36 from whites), by the Sequencher program (Gene Code). Each polymor-
Table 2

Frequency of SNPs in Whites and Japanese

<table>
<thead>
<tr>
<th>SNP Designation</th>
<th>Type</th>
<th>CHIMPANZEE FREQUENCY</th>
<th>Japanese (n = 154 Chromosomes)</th>
<th>White (n = 176 Chromosomes)</th>
<th>Fst</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A-1178G</td>
<td>A</td>
<td>.21</td>
<td>.09</td>
<td>.028</td>
</tr>
<tr>
<td>2</td>
<td>G-1074T</td>
<td>T</td>
<td>.21</td>
<td>.09</td>
<td>.028</td>
</tr>
<tr>
<td>...</td>
<td>T-829A</td>
<td>T</td>
<td>.00</td>
<td>.02</td>
<td>.010</td>
</tr>
<tr>
<td>3</td>
<td>G-792A</td>
<td>A</td>
<td>.21</td>
<td>.09</td>
<td>.028</td>
</tr>
<tr>
<td>4</td>
<td>T-775C</td>
<td>T</td>
<td>.07</td>
<td>.06</td>
<td>.001</td>
</tr>
<tr>
<td>5</td>
<td>C-532T</td>
<td>C</td>
<td>.26</td>
<td>.09</td>
<td>.050</td>
</tr>
<tr>
<td>7</td>
<td>A-20C</td>
<td>C</td>
<td>.24</td>
<td>.16</td>
<td>.010</td>
</tr>
<tr>
<td>8</td>
<td>A-6G</td>
<td>A</td>
<td>.13</td>
<td>.58</td>
<td>.221</td>
</tr>
<tr>
<td>9</td>
<td>C67T</td>
<td>C</td>
<td>.14</td>
<td>.58</td>
<td>.210</td>
</tr>
<tr>
<td>10</td>
<td>C172T</td>
<td>C</td>
<td>.35</td>
<td>.12</td>
<td>.074</td>
</tr>
<tr>
<td>11</td>
<td>G384A</td>
<td>G</td>
<td>.22</td>
<td>.1</td>
<td>.027</td>
</tr>
<tr>
<td>12</td>
<td>G400A</td>
<td>G</td>
<td>.22</td>
<td>.1</td>
<td>.027</td>
</tr>
<tr>
<td>13</td>
<td>G507A</td>
<td>G</td>
<td>.13</td>
<td>.56</td>
<td>.205</td>
</tr>
<tr>
<td>14</td>
<td>A676G</td>
<td>G</td>
<td>.2</td>
<td>.63</td>
<td>.190</td>
</tr>
<tr>
<td>15</td>
<td>A698G</td>
<td>G</td>
<td>.2</td>
<td>.63</td>
<td>.190</td>
</tr>
<tr>
<td>16</td>
<td>A1035G</td>
<td>G</td>
<td>.41</td>
<td>.72</td>
<td>.098</td>
</tr>
<tr>
<td>17</td>
<td>A1164G</td>
<td>G</td>
<td>.38</td>
<td>.83</td>
<td>.212</td>
</tr>
<tr>
<td>18</td>
<td>C2079T</td>
<td>C</td>
<td>.37</td>
<td>.14</td>
<td>.070</td>
</tr>
<tr>
<td>19</td>
<td>G2624A</td>
<td>G</td>
<td>.33</td>
<td>.1</td>
<td>.078</td>
</tr>
<tr>
<td>20</td>
<td>A3189G</td>
<td>A</td>
<td>.35</td>
<td>.07</td>
<td>.118</td>
</tr>
<tr>
<td>21</td>
<td>C3889T(T174M)</td>
<td>C</td>
<td>.16</td>
<td>.14</td>
<td>.001</td>
</tr>
<tr>
<td>...</td>
<td>T3965C(P199P)</td>
<td>T</td>
<td>.00</td>
<td>.01</td>
<td>.005</td>
</tr>
<tr>
<td>22</td>
<td>C4072T(T235M)</td>
<td>C</td>
<td>.12</td>
<td>.56</td>
<td>.216</td>
</tr>
<tr>
<td>23</td>
<td>A5093C</td>
<td>A</td>
<td>.13</td>
<td>.55</td>
<td>.197</td>
</tr>
<tr>
<td>24</td>
<td>C5343T</td>
<td>C</td>
<td>.02</td>
<td>.00</td>
<td>.010</td>
</tr>
<tr>
<td>25</td>
<td>G5556A</td>
<td>G</td>
<td>.13</td>
<td>.56</td>
<td>.205</td>
</tr>
<tr>
<td>26</td>
<td>G5593A</td>
<td>G</td>
<td>.13</td>
<td>.56</td>
<td>.205</td>
</tr>
<tr>
<td>27</td>
<td>A5878C</td>
<td>A</td>
<td>.03</td>
<td>.00</td>
<td>.015</td>
</tr>
<tr>
<td>28</td>
<td>A6066C</td>
<td>C</td>
<td>.44</td>
<td>.78</td>
<td>.121</td>
</tr>
<tr>
<td>29</td>
<td>G6152A</td>
<td>G</td>
<td>.25</td>
<td>.09</td>
<td>.045</td>
</tr>
<tr>
<td>30</td>
<td>C6233T</td>
<td>C</td>
<td>.44</td>
<td>.78</td>
<td>.121</td>
</tr>
<tr>
<td>31</td>
<td>G6309A</td>
<td>G</td>
<td>.34</td>
<td>.65</td>
<td>.096</td>
</tr>
<tr>
<td>32</td>
<td>C6420T</td>
<td>T</td>
<td>.34</td>
<td>.2</td>
<td>.025</td>
</tr>
<tr>
<td>33</td>
<td>C6428G</td>
<td>C</td>
<td>.34</td>
<td>.2</td>
<td>.025</td>
</tr>
<tr>
<td>34</td>
<td>G6442A</td>
<td>G</td>
<td>.08</td>
<td>.04</td>
<td>.007</td>
</tr>
<tr>
<td>35</td>
<td>G7369A</td>
<td>G</td>
<td>.32</td>
<td>.12</td>
<td>.058</td>
</tr>
<tr>
<td>36</td>
<td>C8357T</td>
<td>C</td>
<td>.4</td>
<td>.68</td>
<td>.079</td>
</tr>
<tr>
<td>37</td>
<td>T9597C</td>
<td>T</td>
<td>.33</td>
<td>.12</td>
<td>.063</td>
</tr>
<tr>
<td>38</td>
<td>G9669T</td>
<td>G</td>
<td>.33</td>
<td>.12</td>
<td>.063</td>
</tr>
<tr>
<td>39</td>
<td>A9770G</td>
<td>A</td>
<td>.34</td>
<td>.12</td>
<td>.068</td>
</tr>
<tr>
<td>40</td>
<td>C11535A</td>
<td>C</td>
<td>.05</td>
<td>.32</td>
<td>.121</td>
</tr>
<tr>
<td>41</td>
<td>C11608T</td>
<td>C</td>
<td>.05</td>
<td>.33</td>
<td>.127</td>
</tr>
<tr>
<td>42</td>
<td>G12058A</td>
<td>Deletion</td>
<td>.32</td>
<td>.1</td>
<td>.073</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>Overall</td>
<td>.087</td>
<td></td>
</tr>
</tbody>
</table>

Phenotype has been confirmed by reamplifying and resequencing from the same or the opposite strand. The remainder of the study subjects were sequenced only for the regions in which SNPs were identified in the first set of 72 chromosomes.

Statistical Analysis

In each SNP, the proportion of variation attributable to differences between the Japanese and white populations was estimated by the $F_{st}$ statistic. Haplotype frequencies for multiple loci were estimated by the expectation-maximization (EM) method, by use of the Arlequin program (Schneider et al. 2000), which is available at the ARLEQUIN web site.

Pairwise LD was estimated as $D = x_{ij} - p_1 p_2$, where $x_{ij}$ is the frequency of haplotype $A_i B_j$, and $p_1$ and $p_2$ are the frequencies of alleles $A_i$ and $B_j$ at loci $A$ and $B$, respectively. A standardized LD coefficient, $r$, is given by
Table 3
Nucleotide Diversity

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Mean ± SE Nucleotide Diversity (× 10⁻³)</th>
<th>Japanese (n = 154 Chromosomes)</th>
<th>Whites (n = 174 Chromosomes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>πa</td>
<td>θa b</td>
</tr>
<tr>
<td>Coding (1,458 bp)</td>
<td></td>
<td>3.37 ± 3.22</td>
<td>2.44 ± 1.82</td>
</tr>
<tr>
<td>Noncoding (12,982 bp)</td>
<td></td>
<td>10.50 ± 5.25</td>
<td>5.51 ± 1.53</td>
</tr>
<tr>
<td>Overall (14,400 bp)</td>
<td></td>
<td>9.78 ± 4.88</td>
<td>5.19 ± 1.43</td>
</tr>
</tbody>
</table>

a Average proportion of nucleotide differences between all possible pairs of DNA sequences in sample

b Expected proportion of polymorphic sites, given by \( \frac{S}{n} \), where \( S \) is the number of polymorphic sites in the sequence and \( n \) is the number of sequences.

Figure 2 Pairwise LD between T235M and other SNPs in AGT, evaluated by either \( D' \) (A) or \( r^2 \) (B), in whites and in Japanese. \( D' \) is expressed as an absolute value.
tree for common haplotypes observed in whites and Japanese.

Results

Molecular Variants in AGT

A 14.4-kb genomic region containing the entire AGT was completely sequenced in 18 whites and 18 Japanese. Several known repetitive elements (SINE, LINE, and LTR) and a CA repeat, the microsatellite used for an early linkage study (Jenemaitre et al. 1992), were identified (fig. 1). In total, 44 SNPs (1 polymorphism/327 bp) across the scanned sequence were identified in a total of 72 chromosomes from the 18 whites and the 18 Japanese (fig. 1C). Among these SNPs, transition substitutions were more prevalent (35/44 [79.5%]) than were transversion substitutions (9/44 [20.5%]). Forty-one SNPs were found in noncoding regions, and only three were found in coding regions. Other than the CA repeat, no insertion/deletion polymorphisms were detected.

The 88 white and 77 Japanese subjects included the original 36 individuals who were genotyped for each of the 44 SNPs, by fluorescent sequencing (table 2). Forty SNPs were present in both populations, whereas two SNPs were present only in whites and two other SNPs were present only in Japanese. Fifteen SNPs, including A(-6)G and C4072T (the T235M amino acid polymorphism), showed large frequency differences between whites and Japanese (table 2). The genotype frequencies in our sample fitted Hardy-Weinberg expectations, with remarkable fidelity (data not shown). Chimpanzee sequence, which are useful for estimating the ancestral states of SNPs and haplotypes, were determined at the sites corresponding to human SNPs, by the direct sequencing of products amplifying the BAC DNA containing the chimpanzee AGT sequence (table 2).

The extent of nucleotide diversity in each population is shown in table 3. The average nucleotide diversity, $\pi$, is slightly greater in the Japanese sample (9.78 ± 4.88) than in the white sample (8.36 ± 4.20). The same pattern is observed when $\theta_s$, the expected proportion of polymorphic sites, is measured. Nucleotide diversity is substantially higher in the 13 kb of noncoding DNA than in 1,458 bp of coding sequence. These figures represent slight underestimates because only 72 human chromosomes (36 from Japanese and 36 from whites) were completely sequenced, with the remainder of the sample genotyped only for the 44 polymorphisms defined in the initial sample. Thus, some rare variants are missed, but this would have only a slight effect on the estimates of $\pi$.

LD between T235M and Other SNPs

LD between T235M and other SNPs was studied because of the reported association between the T235 allele and EHT. Figure 2 illustrates substantial differences between $D'$ and $r^2$, in addition to differences between the Japanese and white samples. The $D'$ values are generally much higher than the $r^2$ values, with a large proportion of $D'$ values equal to 1.0 or $-1.0$ (maximum disequilibrium). The percentages of $D'$ values equal to $-1.0$ or 1.0 are 53% in the white sample (412/780 total SNP pairs) and 50% in the Japanese sample (427/861 SNP pairs). The $D'$ values equal to 1.0 were caused by the presence of only three of four possible haplotypes for a pair of loci, which forces $D'$ to its maximum possible value. When LD was evaluated by $r^2$ (fig. 2B), LD with T235M showed several peaks and valleys and no strict correlation with physical distance. In general, LD values were higher in the white sample than in the Japanese sample.

When an arbitrary criterion of $r^2 = 0.5$ was used, eight SNP alleles—A(-6), C67, G507, A676, A698, A5093, G5556, and G5593—were associated with the T235 allele in both populations (table 4). The G6309 and C8357 alleles were associated with T235 only in whites. On the basis of power considerations, Kruglyak (1999) proposed the criterion that $d^2$ values > 0.1 should be considered "useful" levels of LD. Because $r^2$ and $d^2$ are almost perfectly correlated in our sample, we des-

### Table 4

**Physical Distance from and LD with T235M, in Whites and Japanese**

<table>
<thead>
<tr>
<th>Group</th>
<th>RESULTS WHEN DISTANCE FROM T235M IS</th>
<th>0≤1 kb (2 SN Ps)</th>
<th>1=2 kb (6 SN Ps)</th>
<th>2=3 kb (7 SN Ps)</th>
<th>3=4 kb (8 SN Ps)</th>
<th>4=5 kb (8 SN Ps)</th>
<th>5=6 kb (5 SN Ps)</th>
<th>6=7 kb (3 SN Ps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whites</td>
<td>No. (proportion) of SNPs with $r^2 &gt; 0.1$</td>
<td>1 (.50)</td>
<td>6 (1.00)</td>
<td>6 (.86)</td>
<td>8 (1.00)</td>
<td>7 (.88)</td>
<td>5 (1.00)</td>
<td>2 (.67)</td>
</tr>
<tr>
<td></td>
<td>Mean of $r^2$</td>
<td>.102</td>
<td>.588</td>
<td>.29</td>
<td>.45</td>
<td>.39</td>
<td>.159</td>
<td>.24</td>
</tr>
<tr>
<td>Japanese</td>
<td>No. (proportion) of SNPs with $r^2 &gt; 0.1$</td>
<td>0 (.00)</td>
<td>3 (.50)</td>
<td>2 (.29)</td>
<td>4 (.50)</td>
<td>2 (.25)</td>
<td>0 (.00)</td>
<td>2 (.67)</td>
</tr>
<tr>
<td></td>
<td>Mean of $r^2$</td>
<td>.052</td>
<td>.448</td>
<td>.065</td>
<td>.317</td>
<td>.243</td>
<td>.046</td>
<td>.173</td>
</tr>
</tbody>
</table>
ignited $r^2 > 0.1$ as the criterion for useful LD. Table 4 also shows that 35/39 (89%) of the SNPs <7 kb from T235M had an $r^2$ value >0.1 in the white population; in the Japanese population, only 13/39 (33%) of the SNPs met this criterion. As seen in table 5, highly similar values were seen when disequilibrium between each SNP and the A-6G promoter mutation was evaluated. When all the possible pairwise LD values in Japanese, only 13/39 (33%) of the SNPs met this criterion. As expected from the mathematical similarity between $D$' and $r^2$, the pairwise values of these two measures were highly correlated (Pearson's $r = 0.96$). The correlation between $D$' and $D$ was much lower (Pearson's $r = 0.33$), reflecting the large number of $D$' values equal to 1.0 or -1.0.

For assessment of patterns of significant disequilibrium values in the two populations, figure 4 shows pairwise $r^2$ values >0.5 (blackened boxes) and within the range 0.25-0.5 (gray-shaded boxes). The value $r^2 = 0.5$ is equivalent to $\chi^2 = 88$ (P < 10^-19) in 176 white chromosomes and $\chi^2 = 77$ (P < 10^-17) in 154 Japanese chromosomes. The distribution of LD is highly similar in the two populations, and at least five major SNP subgroups with minor changes, defined by having $r^2 = 0.5$, were present (fig. 4, bottom).

Haplotype Analysis

Haplotypes were constructed on the basis of the genotype data from 21 SNPs selected to span most of AGT. Haplotype frequencies were estimated by the EM algorithm, with phase-unknown samples. This procedure has been shown to estimate common haplotype frequencies accurately when the Hardy-Weinberg assumption is fulfilled and when sample sizes are reasonably large (e.g., >100 chromosomes) (Fallin and Schork 2000; Tishkoff et al. 2000a). The haplotypes carrying A(-6) and T235 could be subdivided into five major haplotypes—HA1, HA2, HA3, HA4, and HA5. Only one major haplotype carrying G(-6) and M235, the HG1 haplotype, was present in both populations. Figure 5 shows the haplotypes that were estimated to be present, in two or more copies, in at least one of the populations. Whites and Japanese shared the five frequent haplotypes, even though the frequencies of those haplotypes were quite different between the two populations. In whites, the HG1 haplotype, which is thought to be protective against EHT, had a frequency of 54%. The frequency of this haplotype in the Japanese population was substantially lower, 11%. Haplotype diversity, 2n/(1 -$\sum$x^2)/(2n-1), where x is the frequency of haplotype i and n is sample number, was estimated as 0.684 for the whites and 0.872 for the Japanese.

Recombination Analysis

Evidence of past recombinants in the AGT sequence is given by the DSS (i.e., difference in sum of squares) values, which, in figure 6, are plotted (on the Y-axis) against position in the AGT sequence (on the X-axis). Higher DSS values indicate greater discrepancies between the two trees generated by each half of the sliding
Figure 3  Pairwise LD versus physical distance between all pairwise SNPs, based on the 861 marker pairs in Japanese, evaluated by either $D'$ (A) or $r^2$ (B). Average values of $D'$ and $r^2$ (B) at every 500 bp, in whites and in Japanese, show that LD declines with increasing physical distance between SNP pairs.

window of DNA sequence and thus reflect the likely locations of recombinants. Figure 6 provides evidence for recombinant events at approximately positions 550, 3800, 5600, and 6000 (possible recombinants upstream and downstream of these locations could not be discerned, because of the locations of polymorphisms and because of limitations on the window size). The bootstrap analysis showed that the DSS values at each of these positions differed significantly from 0. These inferred recombinants correspond to blocks of SNPs that are in association with one another, as seen in figures 4 and 5. One block begins with SNP13 (G507A) and ends with SNP17 (A1164G). A second block begins with SNP22 (the T235M polymorphism, C4072T) and ends with SNP28 (A6066C).

Gene Tree for Common Haplotypes Observed in Japanese and Whites

A haplotype tree for the major haplotypes was constructed by use of the ClustalW program (fig. 7). Chimpanzee sequences were used to determine the ancestral haplotype. The HG1 and HA1 haplotypes, the most frequent haplotypes in whites and Japanese, respectively, are remotely related to the chimpanzee sequence.

Relationship between SNP Haplotypes and Microsatellite Marker

The CA repeat, which is located downstream of exon 5, has been identified elsewhere (Jeunemaitre et al. 1992) and was used for linkage studies. The relationship be-
Pairwise LD in AGT, evaluated by $r^2$. LD between all pairs of SNPs (SNP$_i$ and SNP$_j$, where $i$ and $j$ refer to the SNP number shown in table 2) was evaluated by the LD measure $r^2$. Pairwise LD was determined among the 861 marker pairs studied in whites (A) and Japanese (B), and pairs in LD ($r^2 = 0.5$) are shown as blackened boxes. Several subgroups of SNPs in tight LD with each other are shown below. A dot in the center of a square indicates that no data are available, because SNP24 and SNP27 were not observed in whites.
Nakajima et al.: Linkage Disequilibrium in AGT

Figure 5  AGT haplotypes in whites and Japanese. These haplotypes were constructed and the frequencies were estimated by the EM algorithm based on 21 SNPs in AGT. The number of chromosomes analyzed was 176 for whites and 154 for Japanese. Blackened boxes denote the minor allele in Japanese. The chimpanzee sequence is also shown.

between the four most common SNP haplotypes and the microsatellite alleles is shown in figure 8. Although the distribution of CA-repeat alleles varies between whites and Japanese, the association patterns between each SNP haplotype and the microsatellite alleles are very similar in the two populations. The same microsatellite allele is in association with each SNP haplotype in both populations (e.g., microsatellite allele 197 and the HG1 haplotype).

Discussion

The notable successes of LD in localizing the genes responsible for Mendelian disorders (Hästbacka et al. 1994; Feder et al. 1996), combined with the availability of hundreds of thousands of SNPs throughout the genome (Sachidanandam et al. 2001), has sparked a strong interest in the use of LD methods for localizing the genes underlying complex diseases (Risch and Merikangas 1996; Collins et al. 1997; Kruglyak 1999; Jorde 2000; Risch 2000; Jorde et al. 2001; Pritchard and Przeworski 2001; Reich et al. 2001; Schork et al. 2001; Stephens et al. 2001). Many important questions regarding this approach remain unanswered, however. To what extent are LD patterns affected by factors such as chromosome location, isochore structure, and choice of markers? How do evolutionary factors, including natural selection, gene flow, genetic drift, population subdivision, and gene conversion, affect LD? Which types of populations are best suited to LD mapping? Answers to these questions are necessary for the efficient design of LD studies.

Variation in AGT has been shown to correlate with variation in plasma angiotensinogen and with risk of hypertension. Therefore, this gene provides the basis for...
a useful case study of LD patterns in a locus that helps to determine susceptibility to a complex disease. Our results demonstrate that significant LD is found between putative susceptibility alleles in the AGT region and other SNPs. However, the pattern of LD in this region is highly irregular, with some pairs of closely linked SNPs showing little LD. This irregularity has been observed in many previous studies of small genomic regions (MacDonald et al. 1991; Jorde 1995; Jorde et al. 1993, 1994; Nickerson et al. 1998; Taillon-Miller et al. 2000; Abecasis et al. 2001) and is to be expected, because recombination becomes rare relative to other events that can affect LD, such as mutation and gene conversion. Our results show evidence of only a few historical recombinants in this region. This paucity of recombinants helps to explain why $D'$ values are 1.0 for many pairs of polymorphisms: recombination is more likely to generate two new haplotypes from two polymorphic sites, giving rise to a total of four haplotypes. On the other hand, if a new haplotype is generated by mutation, a total of three haplotypes is likely to be seen, and $D'$ for two sites will equal 1.0. The result is that $D'$ is a relatively insensitive measure of LD variation in this small genomic region.

We observed a slightly more regular pattern of LD decline with physical distance when LD values were averaged across 500-bp intervals (fig. 3). This procedure is expected to smooth out some of the variation in LD estimates, and similar results have been obtained in other studies in which LD values are averaged across genomic intervals (Dunning et al. 2000; Abecasis et al. 2001).

Although the average LD values in the present study decline with physical distance, some pairs of SNPs exhibit significant LD at distances of nearly 10 kb. This is consistent with the results of many other empirical studies, some of which detect significant LD at distances up to several hundred kilobases (Jorde et al. 1993, 1994; Peterson et al. 1995; Ajoka et al. 1997; Huttley et al. 1999; Lonjou et al. 1999; Moffatt et al. 2000; Reich et al. 2001; Stephens et al. 2001). These empirical results stand in contrast to a simulation study that predicted little or no useful LD beyond distances of 10 kb (Kruglyak 1999). This study assumed either constant population size or simple exponential growth, both of which are likely to be oversimplifications (Wall and Przeworski 2000). Cyclic bottlenecks and expansions, for example, can lead to higher LD levels (Collins et al. 1999). In addition, the simulation study ignored the potential effects that natural selection has on disease-causing variants. Natural selection limits the length of time during which these variants can persist in populations, reducing the length of time during which LD can dissipate (Terwilliger and Weiss 1998). These and other factors are likely to account for discrepancies between these simulation results and the empirical studies reported thus far.

Comparisons of LD patterns in the Japanese and white populations showed that, although the overall patterns were quite similar, the LD between T235M and
Figure 8  Relationships between four major SNP haplotypes and the microsatellite marker. The distribution of the frequency of individual microsatellite alleles is shown for each of the common SNP haplotypes in AGT. Although the distribution of the CA-repeat alleles (A) in whites is very different from that in Japanese, each SNP haplotype is associated with a specific CA-repeat allele in whites (B) and Japanese (C).
other polymorphisms was substantially greater in the white sample. A total of 89% of the SNPs <7 kb from the EHT-associated T235M polymorphism demonstrated “useful” LD ($r^2 > 0.1$) in the white sample, but this figure was only 33% in the Japanese sample. Thus, the probability of detecting the EHT-associated polymorphism in a genome LD scan would be substantially greater in the white population. The higher level of T235M association is due to an elevated H1 haplotype frequency (54%) in the Utah CEPH sample. This, in turn, may reflect the substantial genetic homogeneity that has been demonstrated in genetic studies of this population (McLellan et al. 1984; O’Brien et al. 1994, 1996). The elevated H1 frequency could also be the result of natural selection affecting the A-6G and/or T235M polymorphisms. Other studies have also demonstrated substantial differences in LD in various populations (Tishkoff et al. 1996, 1998, 2000b; Kidd et al. 1998; Reich et al. 2001), highlighting the effects that population history has on LD patterns.

It is instructive to compare haplotype complexity in AGT with that of the lipoprotein lipase gene, LPL. The AGT region, with an average nucleotide diversity of $p \approx 1/1,000$, is typical of most regions reported thus far (Wall and Przeworski 2000; Jorde et al. 2001; Sachidanandam et al. 2001). LPL has a somewhat higher level of nucleotide diversity, $p = 1/500$, and exhibits a high degree of haplotype complexity in several different populations, with evidence of multiple recombination events (Clark et al. 1998; Nickerson et al. 1998; Templeton et al. 2000). Indeed, haplotype reconstruction showed that, for most (64%) pairs of SNPs in the LPL region, all four haplotypes were present. In contrast, most (50% in the Japanese sample and 53% in the white sample) pairs of SNPs in the AGT region yielded evidence of only three haplotypes, indicating less recombination. Just five major haplotypes (fig. 5) account for 84% of the 176 white chromosomes and for 73% of the 154 Japanese chromosomes. Thus, relatively few SNPs can account for much of the variation in the AGT region, implying that this gene would require a lower SNP density for association detection than would a more complex gene, such as LPL.

Taken together, these results demonstrate that it is not feasible to predict a uniform SNP density for genomewide association studies. The SNP density needed for detection of disease-associated polymorphisms will vary with genomic region, marker type, and choice of population. In addition, the distribution of LD is almost guaranteed to be irregular in relatively small genomic regions, particularly in more-recently-founded populations, which have a relatively brief history of recombination. Additional empirical information about the effects that all these factors have on LD patterns is needed, in order to design efficient association studies.

The haplotype patterns seen in the Japanese and white populations allow some inferences about the history of the EHT-associated AGT polymorphisms. As seen in figure 4, LD and haplotype patterns are quite similar in the two populations, and both share the same major haplotypes (albeit with different frequencies). In addition, the same CA-repeat alleles are found in association with each major haplotype in the two populations. In particular, the M235 allele occurs on the same haplotype background, and this haplotype is quite common in two populations of distinct geographic origin (Japan vs. the northern-European origin of the Utah population). These results, taken together with the fact that the T235M polymorphism is seen in at least some African populations (Corvol and Jeunemaitre 1997), indicate that the polymorphism probably arose before modern humans left Africa and that it was shared by a portion of the population that eventually populated Europe and Asia. Predating the African exodus, the polymorphism is likely to be $\geq 50,000$ years old (Jorde et al. 1998; Hedges 2000; Underhill et al. 2000).

Our results also bear on the question of natural selection for variation in AGT. Notably, the highest $F_{ST}$ values shown in table 2 are those associated with the A-6G promoter polymorphism and the T235M polymorphism, both of which are associated with hypertension. Exceptionally high $F_{ST}$ values are a potential indication of the effects of directional selection (Lewontin and Krakauer 1973; Bowcock et al. 1991; Beaumont and Nichols 1999). An analysis of several nonhuman primate species (chimpanzee, gorilla, orangutan, gibbon, baboon, and macaque) shows that the T235 allele is fixed in these species (Inoue et al. 1997; Dufour et al. 2000). In addition, the A(-6) promoter variant is fixed in the three species examined thus far (chimp, gorilla, and macaque). Thus, the protective M235 and G(-6) variants are likely to have arisen during the course of human evolution. The T235 allele varies widely in frequency, occurring in 35%–45% of whites, 75%–80% of Asians, 75%–80% of African Americans, and $\geq 90\%$ of Africans (Corvol and Jeunemaitre 1997; Staessen et al. 1999). This pattern leads to the hypothesis that the A(-6)/T235 haplotype, associated with higher angiotensinogen expression and greater sodium reabsorption, was adaptive in the tropical, sodium-poor environment of sub-Saharan Africa (Jeanneinaitre et al. 1997) but was selected against (or became selectively neutral) as modern humans radiated out of Africa into other environments. Natural-selection signatures (Kreitman 2000) in AGT should be evaluated in multiple populations, to test this intriguing hypothesis.

**Acknowledgments**

This work was supported in part by a Research for the Future Program Grant, from The Japan Society for the Pro-
motion of Science (to I.I. and M.E.); by Ministry of Public Health and Welfare Research on Human Genome, Tissue Engineering Food Biotechnology (support to I.I.); by National Institutes of Health grants GM-59290 (to L.B.J.) and HL-54471 (to J.-M.L.); and by National Science Foundation grant SBR-9818215 (to L.B.J.). We are grateful to Michael Bamshad and Scott Watkins for comments on the manuscript.

**Electronic-Database Information**

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

ARLEQUIN, http://anthropologie.unige.ch/arlequin (for the Arlequin program)


Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for EHT [MIM 145500] and AGT [MIM 106150])


Topal web site, http://www.rdg.ac.uk/Statistics/software.html (for the Topal 2.0 package)

**References**


