Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels

Xiaofeng Zhu1,*, Laura Fejerman2, Amy Luke1, Adebowale Adeyemo3,4 and Richard S. Cooper1

1Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, 2160 S. First Avenue, Maywood, IL 60153, USA, 2Department of Biological Anthropology, Oxford University, Oxford, UK, 3Department of Pediatrics, University College Hospital, University of Ibadan, Ibadan, Nigeria and 4National Human Genome Research Center, College of Medicine, Howard University, Washington, DC, USA

Received November 18, 2004; Revised December 28, 2004; Accepted January 5, 2005

The most efficient study design to map genes underlying complex traits will be determined by assumptions about whether the genetic effects are likely to be due to relatively few common variants or multiple rare variants. To examine the possibility that rare variants may influence blood pressure, we sequenced a 6.8 kb region of the angiotensinogen (AGT) gene in 29 male Nigerians with high plasma AGT levels and 28 with low levels. The frequency of haplotypes produced from rare variants in the promoter and coding regions was significantly different between the two groups, and it is unlikely that this difference was due to the manner in which the rare variants were selected. Further analysis suggested that most of the haplotypes produced by these rare variants are found on a haplotype background created by three common SNPs. Our study confirms in an additional trait that rare variants can influence the distribution of complex traits; whether these variants can be captured by common SNPs or haplotypes requires further investigation.

INTRODUCTION

The relative frequency of variants that contribute to common human diseases continues to be a subject of intense speculation, in part because the evolutionary context in which these variants arose remains unknown (1). The most common approach to localizing the putative causal variants is based on a comprehensive test for association using all the polymorphisms with a frequency >5–10% (2,3). This approach is based on the assumption that when natural selection is weak, the frequency spectrum of the causative variants is likely to be broad, and many will be common and have a long history in the population (4,5). A recent meta-analysis of genetic epidemiology studies indeed suggests that many common variants contribute to complex disease states, although there are numerous reported associations that are inconsistent (6). Alternatively, other investigators have argued that neutral susceptibility alleles are unlikely to be present at intermediate frequencies and should contribute little to the overall genetic variance for the most plausible range of mutation rates (7–9). Recent studies of HDL-cholesterol as a model trait have in fact found that multiple rare genetic variants in the coding regions of regulatory genes, including ABCA1, APOA1 and LCAT, were significantly over-represented in lower tail of the distribution (10,11), thus confirming, for this trait at least, that many apparent susceptibility loci are rare due to the randomness of the evolutionary processes. These emerging data raise important questions about the efficiency of the ‘haplotype-tagging’ approach to elucidate most of the heritable inter-individual variation.

Rare variants are of recent origin and have a high probability of being lost due to random genetic draft. Potentially, however, variants which confer a selective advantage could be preserved at low frequencies more often than those in other portions of the phenotypic distribution, and this process could account for the observations reported earlier when discordant phenotypes were sampled. On the other hand, as most mutations are deleterious, over-representation of new variants at the bottom of the phenotype distribution could result simply from the fact that they most often confer loss of function.

Enzymatic pathways underlying blood-pressure regulation provide a model system for further investigation of these questions. Plasma angiotensinogen (AGT-P) is the substrate of the
RESULTS

In the present study, we systematically examined the association of the rare haplotypes that resulted from low frequency variants in the promoter and coding regions in AGT with plasma levels of the gene product. We selected 57 males from the highest \((N = 29)\) and from the lowest \((N = 28)\) 20th percentiles of the AGT plasma level distribution in the population sample from Igbo-Ora, Nigeria. We reanalyzed the sequence data and its relation to AGT-P, age, systolic blood pressure (SBP) and diastolic blood pressure (DBP). A total of 93 SNPs were identified across the 6.8 kb sequence. The average nucleotide diversity, \(\pi\), was \(1.77 \times 10^{-3}\), and \(\theta = 2.29 \times 10^{-3}\), which is slightly larger than the previous surveys \((18–20)\). We examined all the SNPs in the promoter region and identified 11 with a minor allele frequency (MAF) <5%. Two SNPs \([C3889T(T174M)\) and \(C4072T(T235M)\)] in exon 2 were also identified with MAF 3.5 and 8.8%, respectively. Table 1 displays the haplotypes of these SNPs, together with variant A-6G, which is in strong linkage disequilibrium (LD) with \(C4072T\).

In the low AGT-P group, we identified 11 haplotypes that were created by the 11 rare SNPs in the promoter region and the exononic SNP \(C3889T\), compared with only three haplotypes in the high group \(P\)-value of Fisher’s exact test is 0.023, Table 2). Because \(C4072T\) has been found to be associated with AGT levels in the literature and is also relatively frequent in Africans, we added this SNP to the analysis. One additional SNP A-6G had MAF <8.8% and was therefore included in the analysis. We then observed 19 rare haplotypes \((H2–H21)\) in the low group, compared with five in the high group \(P\)-value of Fisher’s exact test is 0.0009, Table 2). To estimate the effect of these rare haplotypes on the phenotype, we regressed AGT-P on the haplotypes assuming an additive model (i.e. all the rare haplotypes have the same effect on AGT-P and the effect is additive), with age and BMI as the covariates. Age and BMI were not significant and were then removed. As anticipated, the relationship was highly significant (excluding \(C4072T\): \(\beta = -439.4 \pm 129.2, P = 0.0013\), including \(C4072T\): \(\beta = -472.1 \pm 126.3, P = 0.0004\)). For SBP and DBP, we observed marginal significance when \(C4072T\) was included and after adjustment for AGE and BMI \((SBP: \beta = -8.49 \pm 4.45, P = 0.06, DBP: \beta = -5.0 \pm 3.76, P = 0.19)\); this weak effect could potentially be due to the small sample size.

To test whether the difference in the haplotype frequencies created by the rare SNPs is due to the process we used to select them, we repeated the analyses using the haplotypes created by rare intronic SNPs, which are relatively likely to be neutral. A total of 46 SNPs with an MAF <5% were observed in the introns, as were 50 SNPs with frequencies <9%. We randomly chose 12 SNPs from among the 46 and compared the haplotype frequencies created by these 12 SNPs between the high and the low AGT-P groups. This procedure was repeated 10 000 times, and each time we recorded the \(P\)-value of Fisher’s exact test. We observed a \(P\)-value <0.023 on only nine occasions.

---

**Table 1. Haplotypes formed by rare SNPs in the promoter region and exons**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>−894</th>
<th>−742</th>
<th>−502</th>
<th>−449</th>
<th>−447</th>
<th>−272</th>
<th>−215</th>
<th>−184</th>
<th>−152</th>
<th>−32</th>
<th>−21</th>
<th>−6*</th>
<th>3889</th>
<th>4072</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>53</td>
</tr>
<tr>
<td>H2</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>H3</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>H4</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>H5</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>H6</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>H7</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>H8</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>H9</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>H10</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>H11</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>1</td>
</tr>
</tbody>
</table>

The bold characters represent the rare mutations.

*The frequency of G for SNP −6 is also listed because it is in complete LD with SNP 4072.*
Table 2. Association of rare haplotypes created by rare promoter and coding SNPs in angiotensinogen and the association with angiotensinogen plasma levels using Fisher’s exact test

<table>
<thead>
<tr>
<th></th>
<th>High AGT</th>
<th>Low AGT</th>
<th>Fisher’s exact P</th>
<th>P-value comparing rare SNPs in introns</th>
<th>Regression by additive model β (SD)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4072T excluded</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of H1</td>
<td>55</td>
<td>45</td>
<td>0.023</td>
<td>0.0009</td>
<td>−439.4(129.2)</td>
<td>0.0013</td>
</tr>
<tr>
<td>No. of rare haps</td>
<td>3</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4072T included</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of H1</td>
<td>53</td>
<td>37</td>
<td>0.0009</td>
<td>0.0007</td>
<td>−472.1(126.3)</td>
<td>0.0004</td>
</tr>
<tr>
<td>No. of rare haps</td>
<td>5</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To account for the variation due to the selection of rare SNPs, we compared the same number of randomly selected rare SNPs in introns. P-values were based on 10,000 random draws of the same number of rare SNPs in introns. We also calculated the effect of the rare haplotypes on AGT level using a linear regression assuming additive model.

Similarly, when we selected 14 SNPs from the full set of 50, we observed seven instances of a P-value < 0.0009. These simulation results confirm that the significant difference in rare haplotypes in the promoter regions and the exons is unlikely to be due to random variation in the selection of the rare SNPs.

We then tested whether a subset of common SNPs in AGT could distinguish the haplotypes created by the rare promoter and coding SNPs from the remaining haplotypes. After excluding the rare SNPs, 29 SNPs remained, and their pairwise r² values are presented in Figure 1, as summarized by the software Haploview. We also created a pseudo-SNP (SNP 30 in Fig. 1) which has a value of 1 if one of (H2–H11) is present, otherwise 0. The pseudo-SNP was in lower LD with all the common SNPs, which is not unexpected. By setting the criterion r² > 0.7 in the method proposed by Carlson et al. (19), we identified 16 bins. In each bin, the SNP having the maximum correlation with the pseudo-SNP was selected as the htSNP. To select a subset of htSNPs in AGT to distinguish the haplotypes created by the rare promoter and coding SNPs from the remaining haplotypes, we first defined a dependent variable ‘γ’ which has a value of 1 if one of (H2–H11) is presented, otherwise 0. We then performed stepwise linear regression using ‘γ’ as the dependent variable and htSNPs as independent variables. The final model included SNPs −217, 172, 2186, 5593 and 6420. From these five SNPs, we selected SNPs −217, 172 and 2186, which had the smallest variance, given the rare haplotypes. In other words, there is only one major haplotype which carries most of the rare SNPs in the promoter and coding regions. Among the 24 rare haplotypes produced by the rare promoter and coding SNPs, 19 carried haplotype GCT created by SNPs −217, 172 and 2186, compared with only three in the remaining haplotypes, suggesting that in this instance common SNPs could potentially have captured the variation due to the rare variants (Table 3). The frequencies of the haplotypes that included these three common SNPs were significantly different between the high and the low AGT groups (P = 0.035, Fisher’s exact test).

DISCUSSION

Our study provides evidence that rare sequence variants could have an important influence on variation of blood-pressure regulating phenotypes, in support of the finding by Cohen et al. (10) and Frikke-Schmidt et al. (11) related to HDL-cholesterol. On the basis of extensive simulations, this evidence is unlikely to be due to the process by which we selected the rare SNPs. We did not find significant effects of other covariates on AGT-P, including age (P = 0.72) and BMI (P = 0.38); however, some unmeasured environmental factors could also contribute to the phenotypic variation. At the present time, however, other environmental factors that contribute to the variation of AGT-P have not been described. We identified only two variants in the exons, and our analysis therefore focused on a combination of rare variants in the exons and the promoter region. Molecular cloning and functional analysis have previously suggested a possible role of the AGT promoter in modulating AGT expression (21), with the A(–6) variant conferring 20% greater expression levels (22). Furthermore, evidence of natural selection around the A-6G polymorphism was identified in non-African populations (16). In this study, Tajima’s D (23) did not deviate significantly from the expectation under neutrality (Tajima’s D = −0.739). However, both Fu and Li’s D* and F* (24) rejected the hypothesis of evolutionary neutrality (D* = −3.92, F* = −3.06, P < 0.02). When restricted to the promoter or near SNP C4072T, the hypothesis of evolutionary neutrality was still rejected (D* = −2.56, F* = −2.21, P < 0.05 in promoter; D* = −3.51, F* = −3.42, P < 0.02 near C4072T). These results suggested that haplotypes created by rare variants might be preserved if they had selection value, or they would have disappeared otherwise, consistent with our results.

Our additional finding that most of the rare mutations occurred on haplotype GCT formed by three common SNPs suggests that htSNPs may in some instances capture the haplotype variation due to rare variants, as suggested by Lin et al. (25). However, whether this occurs frequently or has a biological basis remains unknown. Furthermore, our analysis proceeded backwards, aggregating the rare SNPs onto haplotypes; prospectively many possible combinations of SNPs would have to be examined to identify the underlying pattern. Owing to the need to consider a large number of htSNPs, the multiple comparisons adjustment that would be required with this approach would greatly reduce the statistical power of an association study. To generalize, our results suggest that htSNPS, like those that will be provided by
efforts such as the HapMap, might miss a substantial proportion of the variants which underlie complex disease.

**MATERIALS AND METHODS**

**Data collection and genotyping**

All the participants were recruited from Igbo-Ora, a rural community in southwest Nigeria. Study protocols were reviewed and approved by the institutional review boards at University College Hospital, Ibadan, and Loyola University Medical School. Written informed consent was obtained from the participants. The detailed sampling frame was described in our previous studies (26,27). Details concerning sequencing, genotyping and quality control can be found in Fejerman et al. (27). Briefly, the entire AGT gene, spanning 13 kb, was first sequenced in 10 individuals using long-region PCR. There were essentially no variant sites identified in the regions from exon 3 to the 3' region of the gene and therefore this region was excluded for further sequencing. Thus we next sequenced four segments with a total of 6.8 kb distributed over the first 8 kb of the gene in 47 additional individuals. Haplotypes were obtained using the 20 experimentally obtained haplotypes and the software PHASE (28) to improve accuracy.

**Statistical methods**

We compared haplotypes resulting from rare variants in the promoter and coding regions between high and low AGT-P levels by Fisher’s exact test of a 2×2 table. To determine whether the observed significance difference of the rare haplotype frequencies between high and low AGT-P levels was because of the selection of SNPs, we carried out 10 000 simulations by examining the haplotype frequencies resulting from rare intronic SNPs. In each simulation, we randomly selected the same number of intronic SNPs and compared the rare haplotype frequencies between the high and the low AGT-P levels. To obtain an empirical P-value, we counted

---

**Figure 1.** Pairwise $r^2$ values among common SNPs and the pseudo-SNP 30 created by Haploview. The pseudo-SNP 30 is not in strong LD with any of the common SNPs. White: $r^2 = 0$; gray: $0 < r^2 < 1$; black: $r^2 = 1$.

**Table 3.** Number of haplotypes observed that carry haplotypes created by SNPs –217, 172 and 2186

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>No. of haplotypes</th>
<th>Association test, haplotype with plasma level</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCG</td>
<td>4 28</td>
<td>0.035*</td>
</tr>
<tr>
<td>GCT</td>
<td>19 3</td>
<td></td>
</tr>
<tr>
<td>GTT</td>
<td>1 21</td>
<td></td>
</tr>
<tr>
<td>ACT</td>
<td>0 38</td>
<td></td>
</tr>
</tbody>
</table>

*P-value of Fisher’s exact test.
the number of simulations in which the Fisher’s exact test P-value was less than the observed P-value when the rare promoter and coding SNPs were compared.

Pairwise LD between common SNPs was measured by $r^2$ (29). To select htSNPs, we applied one of the accepted methods, [the method of Carlson et al. (19)] which clusters the SNPs into different bins according to the $r^2$ values. We used a criteria $r^2 \geq 0.7$ to define the bins.

**ACKNOWLEDGEMENTS**

We thank Donghui Kan for his assistance in programming. We also thank two anonymous reviews for their helpful comments to further improve the performance of the paper. The work was supported by National Institutes of Health, grant numbers HL074166, HL45508 from NHLBI.

**REFERENCES**


