Published in final edited form as:

Atherosclerosis. 2008 November ; 201(1): 138–147. doi:10.1016/j.atherosclerosis.2008.01.002.

THE APOLIPOPROTEIN(a) GENE: LINKAGE DISEQUILIBRIA AT THREE LOCI DIFFERS IN AFRICAN AMERICANS AND CAUCASIANS

Jill Rubin¹, Han Jo Kim¹, Thomas A. Pearson³, Steve Holleran², Lars Berglund^{1,4,5}, and Rajasekhar Ramakrishnan²

¹Department of Medicine, Columbia University, New York, New York

²Department of Pediatrics, Columbia University, New York, New York

³Department of Community and Preventive Medicine, University of Rochester, Rochester, New York

⁴Department of Medicine, University of California, Davis, Davis, California

⁵VA Northern California Health Care System

Abstract

Lp(a) is an independent, genetically regulated cardiovascular risk factor. Lp(a) plasma levels are largely determined by the apo(a) component, and differ across ethnicity. Although a number of polymorphisms in the apo(a) gene have been identified, apo(a) genetic regulation is not fully understood. To study the relation between apo(a) gene variants, we constructed haplotypes and assessed linkage equilibrium in African Americans and Caucasians for three widely studied apo(a) gene polymorphisms (apo(a) size, +93 C/T and PNR). Apo(a) size allele frequency distributions were different across ethnicity ($p < 0.01$). For African Americans, PNR frequencies were similar across apo(a) sizes, suggesting linkage equilibrium. For Caucasians, the PNR and the PNR-C/T haplotype frequencies differed for large and small apo(a), with the T and PNR 9 alleles associated with large apo(a) size ($p < 0.0002$); also, the PNR 9 allele was more common on a T allele, while PNR 8 was more common on a C allele. On a C allele background, small PNR alleles were more common and the PNR 10 allele less common among African Americans than Caucasians ($p < 0.001$). The ethnic difference in apo(a) size distribution remained controlling for C/T and PNR alleles ($p = 0.023$). In conclusion, allele and haplotype frequencies and the nature of the linkage disequilibrium differed between African Americans and Caucasians at three apo(a) gene polymorphisms.

Keywords

Genetics; African Americans; Genotyping; Polymorphism; Apo(a); PNR; linkage disequilibrium; haplotypes

© 2008 Elsevier Ireland Ltd. All rights reserved.

Correspondence: Lars Berglund, M.D. Department of Medicine University of California, Davis UCD Medical Center, CTSC 2921 Stockton Blvd, Suite 1400 Sacramento, CA 95817 Phone: (916) 703-9120 Fax: (916) 703-9124 e-mail: lars.berglund@ucdmc.ucdavis.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

Apolipoprotein(a), apo(a), is the defining component of lipoprotein(a), Lp(a), an independent, genetically regulated risk factor for cardiovascular disease (1-6). The apo(a) gene, a major predictor of Lp(a) plasma levels, has a limited species distribution and has been detected only in humans, Old World primates and in the hedgehog, although in the latter, apo(a) is molecularly distinct from the first two (1,7,8). Multiple genetic variants have been described for the apo(a) gene, among which a size variation, due to a variable number of repeats coding for a so-called kringle region [kringle 4 (K4)] to a major extent impact plasma Lp(a) levels (9-11). Further, Lp(a) levels vary across African American — Caucasian ethnicity, with levels generally higher among the former (12-17). Although the apo(a) size polymorphism is an important predictor of Lp(a) levels with an inverse relationship between apo(a) size and Lp(a) levels, the latter vary considerably among individuals carrying apo(a) isoforms of the same size, implicating the presence of other predictors (15,18,19). Beyond the apo(a) size variation, two other polymorphisms, an 1 kb upstream pentanucleotide repeat region (PNR) with 5 to 12 TTTTA repeats in the 5'-flanking region of the gene, and a C/T polymorphism at +93 in the promoter region have been studied widely (20-25). We have previously reported on the PNR, C/T, and apo(a) gene size polymorphisms as predictors of allele-specific apo(a) levels and dominance pattern in Caucasians and African Americans (13,26). However, there is a void of information on haplotypes. Although a number of apo(a) polymorphisms have been described, information based on the three studied polymorphisms would provide insight into the relationship between these common gene variants, and provide a basis to deduce haplotype distributions in the two ethnic groups.

METHODS

Study Population

Subjects were recruited from a multiethnic patient population scheduled for diagnostic coronary arteriography either at Harlem Hospital Center in New York City or at the Mary Imogene Bassett Hospital in Cooperstown, NY. Briefly, a total of 648 patients, 401 men and 247 women, ethnically self-identified as 344 Caucasians, 232 African Americans, and 72 Other, were enrolled. Data on apo(a) allele sizes were available in 430 subjects (263 Caucasians, 167 African Americans). The +93 C/T and PNR polymorphisms were determined from DNA samples in 354 subjects (215 Caucasians and 139 African Americans), with complete data available in 264 subjects (160 Caucasian and 104 African American). The design of the Harlem-Bassett study, the recruitment procedure and the clinical characteristics of the subjects have been described previously (13,15,27). The study was approved by the Institutional Review Boards at Harlem Hospital, the Mary Imogene Bassett Hospital, Columbia University College of Physicians and Surgeons, and University of California Davis. Informed consent was obtained from all participants. Details of the determination of the apo(a) size, C/T and PNR repeat polymorphisms are given elsewhere (13,26,28).

Allele Frequency Distributions

The complete apo(a) size genotype distributions for Caucasians and African Americans are given in the online appendix tables. The homozygotes as well as subjects with size difference of 1 K4 repeat are shown in bold and underlined, respectively. For the C/T and PNR polymorphisms, allele frequencies were compared between Caucasians and African Americans by χ^2 analysis. For the gene size polymorphism, the large number of alleles necessitated a modification. First, an overall comparison was done by computing the cumulative frequency distributions for the two ethnic groups using the Kolmogorov-Smirnov test (29). Secondly, allele sizes were grouped into four ranges reflecting the largest ethnic differences in the cumulative frequency distribution and proportions in the ranges compared using χ^2 analysis.

Haplotypes

To estimate the apo(a) size distribution on T and C alleles or on PNR alleles (or PNR distribution on the T allele), it was necessary to determine haplotypes. As haplotypes are ambiguous in subjects heterozygous at two or at all three loci we estimated haplotypes for double and triple heterozygotes as described in detail in the Appendix. Briefly, the primary data are the number of subjects with each multi-locus genotype. To estimate the haplotype frequencies from the primary data, we derived formulas for the probability of observing each multi-locus genotype from haplotype frequencies. Formally, there are two cases: (1) with homozygous genotypes at all loci and (2) with one or more heterozygous loci. For subjects homozygous at either or both of two loci (or at two or three of three loci), haplotypes were determined without the need for estimation. To perform haplotype estimation for subjects with heterozygous loci, we developed software, available online at www.biomath.info/poolfit. For incomplete data, haplotypes over unobserved loci were added up as described in detail in the Appendix. The complexity in enumerating the possible haplotype pairs for each multi-locus genotype, the weighting and the multiple subgroups made this estimation problem intractable with available commercial software packages, prompting us to modify POOLFIT, an existing program which we had developed for fitting lipoprotein tracer data by nonlinear weighted least squares, writing special code for haplotypes, available online at www.biomath.info/poolfit.

As the T allele frequency was very low in African Americans, C/T haplotypes were not calculated for this group. In order to avoid very low haplotype frequencies that would be poorly estimated, grouping of PNR or apo(a) size alleles, respectively, was performed. Estimated haplotype frequencies were used to resolve ambiguities regarding double or triple heterozygotes prior to testing for linkage disequilibrium.

Linkage disequilibrium

Linkage disequilibrium between the C/T and PNR loci in Caucasians was studied by comparing the PNR allele distributions on a C vs T allele background. For subjects homozygous at either or both loci, haplotypes were known unambiguously and the distributions could be determined directly. For example, the 22 Caucasian subjects homozygous for the C allele with a PNR 8/9 genotype contributed 22 PNR 8 and 22 PNR 9 alleles on the C allele background. For double heterozygotes, the haplotype frequencies, estimated as described above, were used: if f_{C8} , f_{C9} , f_{T8} , f_{T9} are the frequencies of the four possible haplotypes for subjects double heterozygous subjects with a C/T and a PNR 8/9 genotype, the fraction of C alleles in the double heterozygotes that carry PNR 8 was calculated as $p = f_{C8} f_{T9} / (f_{C8} f_{T9} + f_{C9} f_{T8})$, and the 31 such subjects were inferred to contribute $31 \times p$. All subjects heterozygous for the C/T and PNR polymorphism were allocated in this fashion.

Statistics

Proportions were compared between groups using χ^2 analysis, and the Fisher exact test where appropriate. All statistical analyses were done using SAS software (SAS Institute, Cary, NC).

RESULTS

Apo(a) Size Allele, C/T and PNR Genotype Frequencies and Distributions

The median apo(a) allele size was 28 K4 repeats for Caucasians and 27 K4 repeats for African Americans, respectively; the lower quartile cut-off was at 23 and 24 K4 repeats for the two groups, and the upper quartile cut-off at 32 and 30 K4 repeats, respectively. To compare the apo(a) size allele frequencies of the two populations, we determined the two cumulative distributions. As shown in Figure 1, the two cumulative allele frequencies were significantly different with a maximum difference of 0.124 at 29 K4 repeats (Kolmogorov-Smirnov statistic

of 1.77, $p=0.004$), indicating that the proportion of African Americans with apo(a) allele sizes <29 K4 repeats was 12.4% higher than the corresponding proportion of Caucasians. As seen in Figure 1, the major differences between the two groups were found in the size ranges around 29 K4 repeats.

Table 1 shows the apo(a) size genotype (Table 1A) and allele frequencies (Table 1B) in the two ethnic populations. To highlight the differences seen in Figure 1, gene sizes were divided into four ranges. As seen in Table 1B, the allele frequencies of the two middle size ranges (25–35 K4 repeats) differed significantly between the two ethnic groups ($p<0.0005$). Thus, 43% of African American alleles were in the 25–29 K4 repeat range compared to 29% of Caucasian alleles, with the proportions reversed in the 30–35 K4 repeat range, confirming the results of the Kolmogorov-Smirnov statistic.

The distributions of genotype combinations for the C/T and PNR polymorphisms are shown in Table 2. As seen in the table, there were no T/T homozygotes and only 7 C/T heterozygotes among African Americans in contrast to 63 T carriers (C/T and T/T) among Caucasian subjects. The total number of C and T alleles were 361 and 69 for Caucasians and 271 and 7 for African Americans, respectively. Further, PNR genotypes with both alleles of smaller sizes (<9) were more common among African Americans compared to Caucasians ($n=82$ vs $n=77$), while PNR genotypes with both alleles representing larger repeat numbers (≥ 9) were more common among Caucasians compared to African Americans ($n=34$ vs $n=11$). For Caucasian PNR alleles, the distribution was 14 for PNR <8; 244 for PNR 8; 93 for PNR 9; 64 for PNR 10 and 15 for PNR 11. Corresponding numbers for African American PNR alleles were 44 for PNR <8; 166 for PNR 8; 47 for PNR 9; 12 for PNR 10; and 9 for PNR 11. Double heterozygotes are shown in bold (<8/8, <8/9, 8/9, 8/10, 9/10 and 9/11 in the C/T columns) and were 7-fold more common among Caucasians compared to African Americans ($n=47$ vs $n=4$, i.e. 21.8% vs 2.9%).

Haplotype Frequencies

Estimated haplotype and allele frequencies for the three apo(a) loci (apo(a) size, PNR and C/T) are given in Table 3 separately for the two ethnic groups. In order to facilitate inter-ethnic comparisons, PNR ranges 5–8 and 10–11 have been combined in both ethnic groups, and apo(a) size was dichotomized at the median (27 K4 repeats) for the combined population. For Caucasians, unlike for African Americans, the distribution of PNR alleles over apo(a) size ranges differed considerably. While the PNR 5–8 range had similar haplotype frequencies for the two apo(a) size ranges (29.5% and 30.6% for 11–27 and 28–45 K4 repeats, respectively) among Caucasians, PNR 9 allele had different frequencies being more common among larger apo(a) sizes (5.0% for smaller apo(a) size and 17.2% for larger apo(a) size). On the other hand, PNR 10–11 had low haplotype frequencies for large apo(a) sizes (3.7%) but higher frequency for smaller apo(a) sizes (14.1%). These findings, coupled with different haplotype frequencies for C and T (e.g., 56.5% for C <8 and 3.6% for T <8, but nearly equal frequency for C 9 and T 9, 11.0% and 11.2%, respectively), suggest linkage disequilibrium among the three loci in Caucasians. Among African Americans, the distribution of PNR alleles over apo(a) sizes was generally more even, suggesting linkage equilibrium between these two loci.

Linkage Disequilibrium

To further investigate linkage disequilibrium, we next analyzed associations between the C/T and PNR polymorphisms in Caucasians, as there were very few African Americans carrying the T allele. Using the estimated haplotype frequencies as described in the Appendix to assign double heterozygotes (shown in bold in Table 2), and the actual haplotypes for the others, we could infer the PNR distribution for Caucasians on either a T or a C background (Table 4). The PNR 9 allele was more common on a T background (67%) compared to a C background (13%), while the PNR 8 allele was more common on a C background (64%) compared to a T

background (19%) ($p<0.0001$), indicating linkage disequilibrium between the two loci for Caucasians.

Next, we compared the PNR distributions between Caucasians and African Americans on a C background. As seen in Table 4, PNR 10 alleles were more common among Caucasian than among African American C alleles (15.4% vs 4.5%, $p=0.0005$). In contrast, small PNR alleles (PNR <8) were more common among African American than among Caucasian C alleles (16.7% vs 3.8%, $p<0.0001$). Thus, after controlling for the C/T polymorphism, the PNR distribution remained significantly different in Caucasians compared to African Americans.

We next explored whether there were any associations between the C/T or PNR polymorphisms and the apo(a) size polymorphism. As seen in Table 5, the C allele was distributed fairly evenly across the two apo(a) size ranges in both ethnic groups (48% and 51% for smaller apo(a) size alleles in African Americans and Caucasians, respectively). In contrast, for Caucasians, the T allele was preferentially associated with larger apo(a) allele sizes; only 8% of T alleles were found in the small apo(a) size range, $p<0.0001$. As suggested by the haplotype frequencies, there was a differential distribution of Caucasian PNR alleles across apo(a) size (Table 5). Compared to the relatively even distribution of the common PNR 8 allele across apo(a) size (42% for small apo(a) size vs 58% (data not shown) for large apo(a) size), smaller apo(a) sizes were less common on the PNR 9 allele (19%) vs 81% (data not shown) for large apo(a) size, $p=0.0002$. On PNR 10, small apo(a) size was more common (83%) vs large apo(a) size (17%, data not shown), $p<0.0001$ (Table 5). For African Americans, the distribution of PNR alleles across apo(a) sizes was in general even, except for PNR 11, where the allele number was lower. The distribution pattern differed significantly between Caucasians and African Americans carrying PNR 9 or 10 ($p=0.005$). The results remained unchanged when restricted to C/C homozygotes in both ethnic groups.

Apo(a) allele size distributions controlling for C/T and PNR alleles

As described above, the largest difference in the cumulative apo(a) size allele frequency between the two ethnic groups was observed at an apo(a) size of 29 K4 repeats (Figure 1). To test whether this difference was modulated by the C/T and PNR mutations, we compared apo(a) size distributions in individuals homozygous for the most prevalent alleles of the two polymorphisms, i.e. both the C and the PNR 8 allele. By building on our finding of the largest ethnic difference in apo(a) allele size frequency at 29 K4 repeats and dichotomizing apo(a) allele sizes at ≤ 29 vs >29 K4 repeats, we confirmed that the apo(a) allele size distribution was significantly different across ethnicity (Table 6). Also for the largest subset, represented by subjects homozygous for the C and PNR 8 alleles, the ethnic difference in apo(a) allele size distribution remained significant (Table 6).

DISCUSSION

Apo(a) is subject to a complex genetic regulation where several polymorphisms impact on plasma Lp(a) levels (9-11). Most studies to date have investigated single polymorphisms and limited information is available on haplotypes. In our study, we characterized the relationship between three widely studied polymorphisms (apo(a) size, +93 C/T and PNR) in Caucasians and African Americans. We have previously reported on the impact of these polymorphisms on total and allele-specific Lp(a) levels (26). Here, we report that allele frequencies at the three loci, haplotype frequencies and nature of the linkage disequilibrium differed across ethnicity.

The genetic variability at the apo(a) locus has attracted attention and many studies have investigated apo(a) size variability (1,30-32). The highly polymorphic nature of the apo(a) gene presents difficulties in analysis of allele frequencies and genotypes. Regarding apo(a) allele size distribution, our results demonstrate a significant difference between African Americans

and Caucasians; African Americans had a significantly higher concentration of apo(a) size genotypes than Caucasians in the 24–30 K4 repeat range. Our finding of an ethnic difference for apo(a) size distribution is in agreement with the results of Kraft et al. (32) and Marcovina et al. (18), but in contrast to those of Gaw et al. (12). Gaw et al. employed a Fisher exact test (using a Markov procedure), which treats allele sizes as distinct categories with no implied ordering. We believe that the allele sizes impose a natural ordering on the categories, allowing us to apply the more powerful Kolmogorov-Smirnov test, also used by Kraft et al. (32). Further, utilizing the allele size frequencies published by Gaw et al, and applying the Kolmogorov-Smirnov test, we found their Caucasian and African American distributions to be different with a P value of 0.022 in contrast to their finding of a nonsignificant $P=0.17$. Thus, apparent discrepancies regarding apo(a) allele frequency distributions between different populations may be due in part to the choice of the statistical procedure.

We observed that small PNR alleles were more common among African Americans, while large PNR alleles were more common among Caucasians. In addition, we demonstrated in Caucasians that the C/T polymorphism was in linkage disequilibrium with both the apo(a) gene size and the PNR polymorphisms; the T allele was more commonly associated with larger apo(a) sizes and the PNR 9 allele. Further, the PNR polymorphism was in linkage disequilibrium with the apo(a) gene size in Caucasians but not in African Americans. Among Caucasians, the PNR 9 allele was more commonly associated with larger apo(a) sizes and the PNR 10 and 11 alleles with smaller apo(a) sizes. Mooser et al reported, using a family study, that carriers of PNR 11 repeats had exclusively small apo(a) size, in their study defined as <24 K4 repeats (23). Our findings in Caucasians also suggest that PNR 10–11 is, albeit not exclusively, associated with small apo(a) sizes. Thus, among Caucasians, at least 3 of eleven PNR 11 alleles had an apo(a) size >24 K4 repeats (one 11/11 homozygote had one apo(a) size allele >24 K4 repeats, and two 11-carriers had both apo(a) size alleles >24 K4 repeats). In contrast, we found that only two of the seven PNR 11 alleles were associated with an apo(a) size <24 K4 repeats among African Americans.

In agreement with the study of Kraft et al (32), we found that the C allele was the most common allele among both Caucasians and African Americans. However, although the T allele frequency among Caucasians in our study (16%) was similar to the one reported by Kraft et al (13%), the African Americans in our study had a lower T allele frequency (2.5%) than observed by them in South African Blacks (9%). The low frequency of the T allele among the African American subjects suggested that it might be explained by Caucasian ancestry. Thus, Parra et al. have estimated Caucasian ancestry in an African American population in New York City to be 20% (the range in samples from 8 other cities in the US was 12% to 23%) (33). Assuming a 15% T allele frequency in Caucasians, a 20% Caucasian ancestry would result in a T allele frequency in African Americans of about 3%, very close to the observed 2.5% in our study. Thus, if we apply the lower and upper ends of the range reported by Parra et al to our results, the expected T frequency in African Americans would be between 1.8% and 3.5%, quite consistent with our findings. These findings suggest a virtual absence of the T allele in the African ancestral areas of the included subjects. However, we cannot exclude a geographical variability in the T allele frequency in different parts of Africa. Of interest, a geographical variability in the apoE4 allele frequency has been documented among African populations (34).

Kraft et al demonstrated that T alleles were associated with larger apo(a) sizes and no T alleles were found among apo(a) sizes <26 K4 repeats (32). We, too, report that the T allele is found mainly with large apo(a) size alleles, though we did observe some T alleles associated with small apo(a) sizes. It is known that the K4 repeat locus has a relatively high mutation rate (35). These results indicate that the C/T mutation could have occurred on a large apo(a) size

allele and subsequently appeared on a range of allele sizes through recombination and/or size mutation.

The present study on haplotypes for the apo(a) gene adds to our knowledge base, and also provides a basis for insight into the relationship between the studied polymorphisms. Our findings on the C/T polymorphism indicates a virtual absence of the T allele among the African ancestry for the African American group in our study, and the differences in PNR haplotypes provides a foundation to explore the etiology of the studied variants. Further, as described in our previous study, the polymorphisms studied impacted differently on allele-specific apo(a) levels in the two ethnic groups (26). As high levels of small size apo(a) is associated with risk for cardiovascular disease, our findings impacts on our ability to more closely evaluate the genetic basis for a risk factor role for Lp(a), and further, potentially may assist in a more targeted intervention.

We recognize that our study has several limitations. Thus, we recruited consecutive patients for cardiac angiography, and the sample therefore does not represent a true population sample. The association between coronary artery disease and small size apo(a), shown by us and others (36-39), might potentially result in a selective allele size distribution among our subjects. However, as previously reported by us, the apo(a) size distribution frequency in the present population was very similar to other studies based on different populations (13), strongly arguing against any selection bias regarding apo(a) sizes. Further, the PNR frequency was similar to that reported in other studies (21-23). As another limitation, due to the complexity of the apo(a) size polymorphism and the need to maintain statistical power, we combined apo(a) size alleles, an approach also used by others. We recognize that there are many ways to pool genotypes and hence there is the possibility of a type 1 error from multiple testing. However, for each analysis, we systematically applied a size grouping that we believe was most relevant, e.g. based on median size or the size with the largest difference in cumulative frequency between the two groups. In spite of these limitations, we believe that our results represent a valid approach to address a complex genetic issue.

In conclusion, our study is one of the first to construct haplotypes based on three apo(a) polymorphisms. The apo(a) gene represents a challenge due to its extensive polymorphism, and the importance of genetic properties in predicting levels of Lp(a), a cardiovascular risk factor. We believe that our novel approach to deduce haplotypes provides insights into the relationship between different genetic variations. Overall, as Lp(a) is largely genetically regulated, a better understanding of the apo(a) gene is needed, and we hope that our study will form a basis for further analysis of the relationship between genetic apo(a) variants in different ethnic groups.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project was supported by grants 49735 (Pearson TA, PI) and 62705 (Berglund L, PI) from the National Heart, Lung and Blood Institute. We thank Mr. Janak Ramakrishnan for assistance in developing the methodology for haplotype estimation. This work was supported in part by the UC Davis Clinical Nutrition Research Unit, NIH #DK35747, and the UC Davis Clinical and Translational Science Center (RR 024146).

APPENDIX

Estimating Haplotype Frequencies from Genotype Data

Existing laboratory methodology leads to a knowledge of the genotype at each locus but does not allow us to determine haplotypes. Further, in some cases there may be missing data. To address these issues, we developed a procedure to estimate haplotype frequencies from genotype data. Considering two loci, if a person is homozygous at one or both, the haplotypes are known uniquely: (C/C, 8/8) must be C8/C8; (C/C, 8/9) must be C8/C9; (C/T, 8/8) must be C8/T8. But the double heterozygote (C/T, 8/9) may be C8/T9 or C9/T8. While an individual subject's haplotypes cannot be determined uniquely, it is possible to use the data from all subjects to estimate the haplotype frequencies for the population under study, as shown below. References are available in population genetics (40-43) to estimate haplotype frequencies, but a mathematical derivation of the necessary results that is easily understood may be of use.

Say there are k loci, with m_1 alleles at locus 1, m_2 alleles at locus 2, etc. The number of possible genotypes at locus j is $g_j = m_j(m_j+1)/2$. For instance, with two alleles a and b , there are three possible genotypes a/a , a/b , b/b ; with three alleles a, b, c , there are six possible genotypes a/a , a/b , a/c , b/b , b/c , c/c . The multilocus genotype for a subject is the collection of the genotypes at all loci. For instance, a 3-locus genotype might be C/T-8/8-Q1/Q3 for the three loci in our study. Please note that in the following, we will use Q to denote size. The number of possible multi-locus genotypes is the product of the possible genotypes at all the loci. For example, Caucasians with 2 alleles at C/T, 3 at PNR, and 4 at size, have $3 \times 6 \times 10 = 180$ possible 3-locus genotypes.

The observations consist of the multi-locus genotype for each subject. Thus the primary data are the number of subjects with each multi-locus genotype. Haplotypes specify the alleles that are on the same chromosome. In our study with three loci, C9Q3 specifies that a chromosome contains the C allele, PNR 9, and size in the Q3 range. The goal here is to estimate the haplotype frequencies f from the primary data. To do so, we derive formulas for the probability of observing each multi-locus genotype from haplotype frequencies. A necessary assumption is Hardy-Weinberg equilibrium so that the probability of observing a haplotype in a subject is independent of the other haplotype. This assumption is satisfied in our study since HWE was seen in C/T and PNR loci and also for the size locus when sizes were grouped into ranges.

A multi-locus genotype can arise in one or more ways. For instance, C/C-8/8-Q1/Q1 can arise in only one way — two identical haplotypes C8Q1; the probability of observing this 3-locus genotype is simply

$$p(C/C-8/8-Q1/Q1) = f_{C8Q1}^2 \quad (1)$$

The 3-locus genotype C/C-8/8-Q1/Q3 can arise when the haplotype in the first chromosome is C8Q1 and that on the second chromosome is C8Q3, or when the first chromosome has C8Q3 and the second C8Q1. There are two ways but they have the same probability, so

$$p(C/C-8/8-Q1/Q3) = 2 f_{C8Q1} f_{C8Q3} \quad (2)$$

A multi-locus genotype with more than one heterozygous locus has more possibilities. For instance, C/C-8/9-Q1/Q3 can arise from C8Q1/C9Q3, C9Q3/C8Q1, C8Q3/C9Q1, or C9Q1/C8Q3 —

$$p(C/C-8/9-Q1/Q3) = 2(f_{C8Q1} f_{C9Q3} + f_{C8Q3} f_{C9Q1}) \quad (3)$$

As the number of heterozygous loci increases, so does the number of ways the multi-locus genotype can arise — it is 2^h where h is the number of heterozygous loci. For instance, C/T-8/9-Q1/Q3 with three heterozygous loci, can arise in one of eight ways - C8Q1/T9Q3, T9Q3/C8Q1, C8Q3/T9Q1, T9Q1/C8Q3, C9Q1/T8Q3, T8Q3/C9Q1, C9Q3/T8Q1, or T8Q1/C9Q3, with the probability

$$p(C/T-8/9-Q1/Q3) = 2(f_{C8Q1} f_{T9Q3} + f_{C8Q3} f_{T9Q1} + f_{C9Q1} f_{T8Q3} + f_{C9Q3} f_{T8Q1}) \quad (4)$$

More formally, there are two cases: (1) with homozygous genotypes at all loci — i_1/i_1 at the first locus, j_1/j_1 at the second locus, etc.,

$$p(i_1/i_1-j_1/j_1-\dots) = f_{i_1j_1}^2 \dots \quad (5)$$

(2) with one or more heterozygous loci, and genotypes i_1/i_2 at the first locus, j_1/j_2 at the second locus, etc.,

$$p(i_1/i_2-j_1/j_2-\dots) = 2 \sum f_{ij} \dots f_{i'j'} \dots \quad (6)$$

where $i=i_1$, $i'=i_2$, or $i=i_2$, $i'=i_1$, etc., and the summation is over the 2^{h-1} ways of constructing the haplotypes.

The minimization criterion is the logarithm of the EM criterion (43):

$$S = \sum n_g \log p_g \quad (7)$$

where n_g is the number of subjects observed with the multi-locus genotype g , n is the total number of subjects, p_g is the probability of the multi-locus genotype g , and the summation is over all possible multi-locus genotypes. During the iterative process, initial convergence was found to be better with the least squares approximation to the criterion:

$$S = \sum (n_g - n p_g)^2 / n p_g (1 - p_g) \quad (8)$$

Incomplete Data

Not all subjects have genotype information at all loci. There may be technical or other reasons. In our study, 160 Caucasians had all 3 loci genotyped, 55 had only C/T and PNR genotyped, and 103 had only size genotyped. These subgroups are handled by constructing an S for each subgroup and adding them up. In our study,

$$S = S_{(all\ 3)} + S_{(C/T\ PNR)} + S_{(size)} \quad (9)$$

The n to be used for each S is just the number of subjects in that subgroup. Also, for the incomplete subgroups, the haplotypes are obtained by adding up the haplotypes over the unobserved loci. For instance, for C/T PNR (no size genotype information),

$$p_{C8} = p_{C8Q1} + p_{C8Q2} + p_{C8Q3} + p_{C8Q4} \quad (10)$$

and similarly for the other two-locus haplotypes. For size alone (no C/T or PNR genotype information),

$$p_{Q1} = p_{C<8Q1} + p_{T<8Q1} + p_{C9Q1} + p_{T9Q1} + p_{C>9Q1} + p_{T>9Q1} \quad (11)$$

and similarly for p_{Q2} , etc.

The complexity in enumerating the possible haplotype pairs for each multi-locus genotype, the weighting and the multiple subgroups made this estimation problem intractable with standard packages such as SAS. We modified POOLFIT, which we had developed for fitting lipoprotein tracer data (44) by nonlinear weighted least squares, writing special code for haplotypes. The haplotype-estimating version of POOLFIT is available online at www.biomath.info/poolfit.

REFERENCES

1. Utermann G. The mysteries of lipoprotein(a). *Science* (Wash DC) 1998;246:904–10.
2. Stein JH, Rosenson RS. Lipoprotein Lp(a) excess and coronary heart disease. *Arch. Intern. Med* 1997;157:1170–6. [PubMed: 9183227]
3. Danesh J, Collins R, Peto R. Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies. *Circulation* 2000;102:1082–5. [PubMed: 10973834]
4. Rhoads GG, Dahlén G, Berg K, Morton NE, Dannenberg AL. Lp(a) lipoprotein as a risk factor for myocardial infarction. *JAMA* 1986;256:2540–4. [PubMed: 2945939]
5. Bostom AG, Gagnon DR, Cupples LA, Wilson PWF, Jenner JL, Ordovas JM, Schaefer EJ, Castelli WP. A prospective investigation of elevated lipoprotein(a) detected by electrophoresis and cardiovascular disease in women. The Framingham Heart Study. *Circulation* 1994;90:1688–95. [PubMed: 7923652]
6. Schaefer EJ, Lamon-Fava S, Jenner JL, McNamara JR, Ordovas JM, Davis CE, Abolafia JM, Lippel K, Levy RI. Lipoprotein(a) and risk of coronary heart disease in men: the lipid research clinics coronary primary prevention trial. *JAMA* 1994;271:999–1003. [PubMed: 8139085]
7. Lawn RM. How often has Lp(a) evolved? *Clin. Genet* 1996;49:167–74. [PubMed: 8828980]
8. Lawn RM, Boonmark NW, Schwartz K, Lindahl GE, Wade DP, Byrne CD, Fong KJ, Meer K, Patthy L. The recurring evolution of lipoprotein(a). Insights from cloning of hedgehog apolipoprotein(a). *J. Biol. Chem* 1995;270:24004–9. [PubMed: 7592597]
9. Gaubatz JW, Ghanem KI, Guevara J Jr, Nava ML, Patsch W, Morrisett JD. Polymorphic forms of human apolipoprotein(a): Inheritance and relationship of their molecular weights to plasma levels of lipoprotein(a). *J. Lipid Res* 1990;31:603–13. [PubMed: 2141054]
10. Gavish D, Azrolan N, Breslow J. Plasma Lp(a) concentration is inversely correlated with the ratio of Kringle IV/Kringle V encoding domains in the apo(a) gene. *J. Clin. Invest* 1989;84:2021–7. [PubMed: 2556454]
11. Sandholzer C, Hallman DM, Saha N, Sigurdsson G, Lackner C, Czsászár A, Boerwinkle E, Utermann G. Effects of the apolipoprotein(a) size polymorphism on the lipoprotein(a) concentration in 7 ethnic groups. *Hum. Genet* 1991;86:607–14. [PubMed: 2026424]
12. Gaw A, Boerwinkle E, Cohen JC, Hobbs HH. Comparative analysis of the apo(a) gene, apo(a) glycoprotein, and plasma concentrations of Lp(a) in three ethnic groups: evidence for no common “null” allele at the apo(a) locus. *J. Clin. Invest* 1994;93:2526–34. [PubMed: 8200989]

13. Rubin J, Paultre F, Tuck CH, Holleran S, Reed RG, Pearson TA, Thomas CM, Ramakrishnan R, Berglund L. Apolipoprotein(a) genotype influences isoform dominance pattern differently in African Americans and Caucasians. *J. Lipid Res* 2002;43:234–44. [PubMed: 11861665]
14. Marcovina SM, Albers JJ, Jacobs DR Jr, Perkins LL, Lewis CE, Howard BV, Savage P. Lipoprotein (a) concentrations and apolipoprotein(a) phenotypes in Caucasians and African Americans: the CARDIA study. *Arterioscler. Thromb* 1993;13:1037–45. [PubMed: 8318505]
15. Paultre F, Pearson TA, Weil HFC, Tuck CH, Myerson M, Rubin J, Francis CK, Marx H, Philbin E, Reed RG, Berglund L. High levels of lipoprotein(a) carrying a small apolipoprotein(a) isoform is associated with coronary artery disease in both African American and Caucasian men. *Arterioscler. Thromb. Vasc. Biol* 2000;20:2619–24. [PubMed: 11116062]
16. Mooser V, Scheer D, Marcovina SM, Wang J, Guerra JR, Cohen J, Hobbs HH. The apo(a) gene is the major determinant of variation in plasma Lp(a) levels in African Americans. *Am. J. Hum. Genet* 1997;61:402–17. [PubMed: 9311746]
17. Parra HJ, Luyeye I, Bouramou C, Demarquilly C, Fruchart JC. Black-white differences in serum Lp (a) lipoprotein levels. *Clin. Chim. Acta* 1987;167:27–31. [PubMed: 3665103]
18. Marcovina SM, Albers JJ, Wijsman E, Zhang Z, Chapman NH, Kennedy H. Differences in Lp(a) concentrations and apo(a) polymorphs between black and white Americans. *J. Lipid Res* 1996;37:2569–85. [PubMed: 9017509]
19. Geethanjali FS, Luthra K, Lingenhel A, Kanagasabapathy AS, Jacob J, Srivastava LM, Vasisht S, Kraft HG, Utermann G. Analysis of the apo(a) size polymorphism in Asian Indian populations: association with Lp(a) concentration and coronary heart disease. *Atherosclerosis* 2003;169:121–30. [PubMed: 12860258]
20. Zysow BR, Lindahl GE, Wade DP, Knight BL, Lawn RM. C/T polymorphism in the 5' untranslated region of the apolipoprotein(a) gene introduces an upstream ATG and reduces in vitro translation. *Arterioscler. Thromb. Vasc. Biol* 1995;15:58–64. [PubMed: 7749816]
21. Trommsdorff M, Köchl S, Lingenhel A, Kronenberg F, Delport R, Vermaak H, Lemming L, Klausen IC, Faergeman O, Utermann G, Kraft HG. A pentanucleotide repeat polymorphism in the 5' control region of the apolipoprotein(a) gene is associated with lipoprotein(a) plasma concentrations in Caucasians. *J. Clin. Invest* 1995;96:150–7. [PubMed: 7615785]
22. Valenti K, Aveyrier E, Leauté S, Laporte F, Hadjian AJ. Contribution of apolipoprotein(a) size, pentanucleotide TTTTA repeat and C:T (+93) polymorphisms of the apo(a) gene to regulation of lipoprotein(a) plasma levels in a population of young European Caucasians. *Atherosclerosis* 1999;147:17–24. [PubMed: 10525120]
23. Mooser V, Mancine FP, Bopp S, Peto-Schramm A, Guerra R, Boerwinkle E, Müller HJ, Hobbs HH. Sequence polymorphisms in the apo(a) gene associated with specific levels of Lp(a) in plasma. *Hum. Mol. Genet* 1995;4:173–81. [PubMed: 7757064]
24. Kraft HG, Windeger M, Menzel HJ, Utermann G. Significant impact of the +93 C:T polymorphism in the apolipoprotein(a) gene on Lp(a) concentrations in Africans but not in Caucasians: confounding effect of linkage disequilibrium. *Hum. Mol. Genet* 1998;7:257–64. [PubMed: 9425232]
25. Kalina A, Csaszar A, Fust G, Nagy B, Szalai C, Karadi I, Duba J, Prohaszka Z, Horvath L, Dieplinger H. The association of serum lipoprotein(a) levels, apolipoprotein(a) size and (TTTTA)(n) polymorphism with coronary heart disease. *Clin. Chim. Acta* 2001;309:45–51. [PubMed: 11408005]
26. Rubin J, Kim HJ, Pearson TA, Holleran S, Ramakrishnan R, Berglund L. The APO[a] size and PNR polymorphisms explain African American-Caucasian differences in allele-specific apo[a] levels for small but not large apo[a]. *J Lipid Res* 2006;47:982–9. [PubMed: 16495513]
27. Jiang XJ, Paultre F, Pearson TA, Reed RG, Francis CK, Lin M, Berglund L, Tall AR. Plasma sphingomyelin level as a risk factor for coronary artery disease. *Arterioscler. Thromb. Vasc. Biol* 2000;20:2614–8. [PubMed: 11116061]
28. Rubin J, Pearson TA, Reed RG, Berglund L. Fluorescence-based, nonradioactive method for efficient detection of the pentanucleotide repeat (TTTTA)_n polymorphism in the apolipoprotein(a) gene. *Clin. Chem* 2001;47:1758–62. [PubMed: 11568083]
29. Hollander, M.; Wolfe, DA. *Nonparametric Statistical Methods*. John Wiley; New York: 1973.
30. Utermann G. Genetic architecture and evolution of the lipoprotein(a) trait. *Curr. Opin. Lipidol* 1999;133–41. [PubMed: 10327281]

31. Marcovina, SM.; Koschinsky, ML. Lipoprotein (a): Structure, measurement, and clinical significance.. In: Rifai, N.; Warnick, GR.; Dominiczak, MH., editors. Handbook of Lipoprotein Testing. AACC Press; 1997. p. 283-313.
32. Kraft HG, Lingenhel A, Pang RW, Delport R, Trommsdorff M, Vermaak H, Janus ED, Utermann G. Frequency distributions of apolipoprotein(a) kringle IV repeat alleles and their effects on lipoprotein (a) levels in Caucasian, Asian, and African populations: the distribution of null alleles is non-random. *Eur. J. Hum. Genet* 1996;4:74–87. [PubMed: 8744025]
33. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD. Estimating African American admixture proportions by use of population-specific alleles. *Am J Hum Genet* 1998;63:1839–51. [PubMed: 9837836]
34. Wozniak MA, Faragher EB, Todd JA, Koram KA, Riley EM, Itzhaki RF. Does apolipoprotein E polymorphism influence susceptibility to malaria? *J. Med. Genet* 2003;40:348–51. [PubMed: 12746397]
35. Boerwinkle E, Leffert CC, Lin J, Lackner C, Chiesa G, Hobbs HH. Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *J. Clin. Invest* 1992;90:52–60. [PubMed: 1386087]
36. Sandholzer C, Saha N, Kark JD, Rees A, Jaross W, Dieplinger H, Hoppichler F, Boerwinkle E, Utermann G. Apo(a) isoforms predict risk for coronary heart disease. A study in six populations. *Arterioscler Thromb* 1992;12:1214–26. [PubMed: 1390593]
37. Longenecker JC, Klag MJ, Marcovina SM, Powe NR, Fink NE, Giaculli F, Coresh J. Small apolipoprotein(a) size predicts mortality in end-stage renal disease: The CHOICE study. *Circulation* 2002;106:2812–8. [PubMed: 12451008]
38. Wu HD, Berglund L, Dimayuga C, Jones J, Sciacca RR, Di Tullio M, Homma S. High lipoprotein (a) levels and small apolipoprotein(a) sizes are associated with endothelial dysfunction in a multiethnic cohort. *J Am Coll Cardiol* 2004;43:1828–33. [PubMed: 15145108]
39. Kronenberg F, Kronenberg MF, Kiechl S, Trenkwalder E, Santer P, Oberhollenzer F, Egger G, Utermann G, Willeit J. Role of lipoprotein(a) and apolipoprotein(a) phenotype in atherogenesis: prospective results from the Bruneck study. *Circulation* 1999;100:1154–60. [PubMed: 10484534]
40. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol* 1995;12:921–7. [PubMed: 7476138]
41. Schipper RF, D'Amato J, de Lange P, Schreuder GMTh, van Rood JJ, M. Oudshoorn M. Validation of haplotype frequency estimation methods. *Hum. Immunol* 1998;59:518–23. [PubMed: 9712357]
42. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet* 2000;67:947–59. [PubMed: 10954684]
43. Clark VJ, Metheny N, Dean M, Peterson RJ. Statistical estimation and pedigree analysis of CCR2-CCR5 haplotypes. *Hum. Genet* 2001;108:484–493. [PubMed: 11499673]
44. Berglund L, Witztum JL, Galeano NF, Khouw AS, Ginsberg HN, Ramakrishnan R. Three-fold effect of lovastatin treatment on low density lipoprotein metabolism in subjects with hyperlipidemia: increase in receptor activity, decrease in apoB production, and decrease in particle affinity for the receptor. Results from a novel triple-tracer approach. *J. Lipid Res* 1998;39:913–24. [PubMed: 9555954]

Figure 1.

Kolmogorov-Smirnov plot of accumulated apo(a) size frequencies in Caucasians and African Americans. The vertical dotted line indicate the K4 size (29 K4) with the largest difference between African Americans and Caucasians, and the horizontal lines indicate the magnitude of this difference.

TABLE 1

Distribution of apo(a) size genotypes and alleles in apo(a) size groups for 167 African Americans and 263 Caucasians

A. Distribution of apo(a) size genotypes					
Apo(a) size genotypes	263 Caucasians			167 African Americans	
	10–24	25–29	30–35	10–24	25–29
10–24	27	57	46	12	17
25–29		16	58	8	24
30–35			24	14	
>35				1	

B. Distribution of apo(a) size alleles					
Apo(a) size alleles	526 Caucasian alleles			334 African American alleles	
	10–24	25–29	30–35	10–24	25–29
Number of Alleles (n)	169	155	166	36	105
Allele frequency (%)	32	29*	32*	7	31

Apo(a) sizes are given as K4 repeat number. In panel A, genotypes are given as combination of size groups for the two alleles, and numbers represent individuals with the respective genotypes. In panel B, numbers represent size groups for apo(a) alleles.

* p=0.0001, proportion of alleles among Caucasians compared to African Americans in the 25–29 and 30–35 K4 repeat allele size ranges.

Distribution of C/T and PNR genotypes in 139 African Americans and 215 Caucasians

TABLE 2

	Caucasians				African Americans			
	C/C	C/T	T/T	Total for PNR genotypes	C/C	C/T	T/T	Total for PNR genotypes
<8/8	3	0	0	3	5	0	0	5
<8/8	5	1	0	6	28	0	0	28
<8/9	1	1	0	2	4	0	0	4
<8/10	0	0	0	0	2	0	0	2
<8/11	0	0	0	0	0	0	0	0
8/8	62	5	1	68	46	3	0	49
8/9	22	31	3	56	25	4	0	29
8/10	35	6	0	41	7	0	0	7
8/11	5	0	0	5	4	0	0	4
9/9	6	3	1	10	6	0	0	6
9/10	2	7	1	10	1	0	0	1
9/11	4	1	0	5	1	0	0	1
10/10	5	1	0	6	1	0	0	1
10/11	1	0	0	1	0	0	0	0
11/11	1	1	0	2	2	0	0	2
Total for C/T genotypes	152	57	6	215	132	7	0	139

Numbers represent individuals with the respective PNR and C/T genotypes and numbers given in bold represent individuals heterozygous for both the C/T and PNR polymorphism.

TABLE 3

A. Relative haplotype and allele frequency estimates at three apo(a) loci (apo(a) size, PNR and C/T) for Caucasians				
PNR	C/T	Apo(a) size (K 4 repeats)		
		11-27	28-45	11-45
5-8	C	28.5	28.0	56.5
	T	0.9	2.6	3.6
	C+T	29.5	30.6	60.1
9	C	3.9	7.0	11.0
	T	1.1	10.1	11.2
	C+T	5.0	17.2	22.2
10-11	C	13.3	2.2	15.6
	T	0.8	1.5	2.1
	C+T	14.1	3.7	17.7
All (5-11)	C	45.8	37.3	83.1
	T	2.7	14.2	16.9
	C+T	48.6	51.4	100.0
B. Relative haplotype and allele frequency estimates at two apo(a) loci (apo(a) size and PNR) for African Americans				
PNR	Apo(a) size (K4 repeats)			
	11-27	28-45	11-45	
5-8	42.0	33.3	75.4	
9	10.1	6.9	17.1	
10-11	3.1	4.4	7.5	
All (5-11)	55.5	44.5	100.0	

Numbers represent frequencies given as percentages of the total number of subjects (160 Caucasians and 104 African Americans).

Relative distribution of PNR alleles on C and T alleles in Caucasians and African Americans

TABLE 4

Alleles	PNR <8	PNR 8	PNR 9	PNR 10	PNR 11
Caucasians T allele (n=69)	0.5	19.0	66.9*	12.0	1.6
Caucasians C allele (n=361)	3.8***	64.0	13.0	15.4**	3.8
African Americans C allele (n=264)	16.7***	59.1	16.3	4.5**	3.4

All frequencies are given as percentages.

* p<0.0001, frequency of PNR 8 and 9 on T vs. C allele in Caucasians

** p=0.0005, frequency of PNR 10 vs. 8 on C allele in Caucasians and AA.

*** p<0.0001, frequency of PNR <8 vs 8 on C allele in Caucasians and AA.

TABLE 5

Distribution of small apo(a) gene size on C/T and PNR alleles (% small apo(a) defined as % in apo(a) size range of K4 repeat 11–27)

	Alleles	C	T	PNR <8	PNR 8	PNR 9	PNR 10	PNR 11	All
Caucasians	n	271	49	12	176	72	49	11	320
	% small apo(a)	51%	8%	38%	42%	19% ^{*,**,*}	83% ^{*,**,*}	77% ^{*,**,*}	44%
African Americans	n	196		34	128	29	10	7	208
	% small apo(a)	48%	n.d.	43%	49%	49% [#]	37% [#]	0%	46%

Numbers given represent alleles from 160 Caucasians and 104 African Americans.

n.d. = not determined

, *, p=0.04, p=0.0002, p<0.0001, % small apo(a) in PNR 11, 9 or 10 alleles respectively, vs % small apo(a) in PNR 8 allele for Caucasians
p=0.005, % small apo(a) in Caucasians vs. African Americans for PNR 9 and 10 alleles, respectively
§ p<0.0001, % small apo(a) in C vs. T alleles for Caucasians

TABLE 6

Apo(a) allele size distribution by ethnicity in all subjects and in subjects homozygous for the C and PNR 8 alleles

	<29 K4 repeats			≥29 K4 repeats			Total	chi-square	p-value
	Caucasians	African Americans	Total	Caucasians	African Americans	Total			
All	203	154	357	117	54	171	528	0.014	
C/C.8/8	57	50	107	19	38	57	164	0.023	

Numbers represent alleles for 160 Caucasians and 104 African Americans (upper panel) and 44 Caucasians and 38 African Americans homozygous for the C and PNR 8 alleles (lower panel).