Published in final edited form as:

Genomics. 2004 May ; 83(5): 912–923. doi:10.1016/j.ygeno.2003.11.016.

Haplotype analysis of the apolipoprotein gene cluster on human chromosome 11

Michael Olivier^{a,*}, Xujing Wang^b, Regina Cole^a, Brian Gau^a, Jessica Kim^a, Edward M. Rubin^{c,d}, and Len A. Pennacchio^{c,d}

^a Human and Molecular Genetics Center, Department of Physiology, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

^bMcGee Center for Juvenile Diabetes, Medical College of Wisconsin, Milwaukee, WI 53226, USA

^c Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

^dJoint Genome Institute, Walnut Creek, CA, USA

Abstract

Members of the apolipoprotein gene cluster (*APOA1/C3/A4/A5*) on human chromosome 11q23 play an important role in lipid metabolism. Polymorphisms in both *APOA5* and *APOC3* are strongly associated with plasma triglyceride concentrations. The close genomic locations of these two genes as well as their functional similarity have hindered efforts to define whether each gene independently influences human triglyceride concentrations. In this study, we examined the linkage disequilibrium and haplotype structure of 49 SNPs in a 150-kb region spanning the gene cluster. We identified a total of five common *APOA5* haplotypes with a frequency of greater than 8% in samples of northern European origin. The *APOA5* haplotype block did not extend past the 7 SNPs in the gene and was separated from the other apolipoprotein gene in the cluster by a region of significantly increased recombination. Furthermore, one previously identified triglyceride risk haplotype of *APOA5* (*APOA5**3) showed no association with three *APOC3* SNPs previously associated with triglyceride concentrations, in contrast to the other risk haplotype (*APOA5**2), which was associated with all three minor *APOC3* SNP alleles. These results highlight the complex genetic relationship between *APOA5* and *APOC3* and support the notion that *APOA5* represents an independent risk gene affecting plasma triglyceride concentrations in humans.

Keywords

Single nucleotide polymorphism; Apolipoprotein A5; Haplotype; Linkage disequilibrium; Recombination; Four-gamete test

The apolipoprotein gene cluster on human chromosome 11q23 contains four apolipoprotein genes (*APOA1/C3/A4/A5*) in a genomic interval of approximately 60 kb [1]. Three of these genes (*APOA1/C3/A4*) have been well described, and each plays an important role in lipid metabolism in humans and mice. For example, mice lacking *apoA1* have significantly reduced plasma high-density lipoprotein (HDL) cholesterol levels [2]. In contrast, mice lacking *apoC3* show lower concentration of plasma triglycerides compared to control littermates [3].

In humans, analyses of genetic sequence variants around these three genes have revealed polymorphisms associated with plasma lipid levels (for a review, see [4]). Sequence variants

*Corresponding author. Fax: +1-414-456-6516. molivier@mcw.edu (M. Olivier).

in *APOA1* primarily affect HDL-cholesterol concentrations, while variation in *APOC3* is primarily associated with altered plasma triglyceride concentrations. In *APOC3*, two rare alleles in the promoter region ($-482C \rightarrow T$ and $-455T \rightarrow C$) and a minor allele in the 3'UTR (*SstI* polymorphism, $3238G \rightarrow C$) have repeatedly been associated with elevated plasma triglyceride concentrations in several human populations [5–15]. However, the lack of strong functional data at least for the *SstI* polymorphism raises questions whether the association seen in humans is due to these sequence variants in *APOC3* directly or due to other functional variants in *APOC3* or possibly in one of the neighboring apolipoprotein genes.

Recently, we identified a fourth member (*APOA5*) of the chromosome 11 apolipoprotein gene cluster, located approximately 27 kb distal to *APOA4* and 37 kb from *APOC3* [1]. Similar to *APOC3*, *APOA5* has been shown to be involved in plasma triglyceride level regulation in both humans and mice. Mice overexpressing human *APOA5* have decreased plasma triglyceride concentrations, while mice lacking *apoA5* have increased plasma concentrations of triglycerides [1]. Similarly, initial studies in humans showed that three single-nucleotide polymorphisms (SNPs 1–3) in and around *APOA5* were significantly associated with altered plasma triglyceride concentrations in two human populations [1]. A fourth SNP (called SNP 4), located between *APOA5* and the proximal *APOA4* gene, was not associated with any plasma lipid measures. Additional association analysis of the *APOC3 SstI* polymorphism in this population revealed no association with triglycerides. This results suggested that the association found with *APOA5* sequence variants was independent of the previously reported effect of the *APOC3 SstI* polymorphism on plasma triglyceride concentrations. The association of *APOA5* has been confirmed in other studies [16–19]. In our subsequent analysis of multiple ethnic groups, we identified additional sequence variants in and around *APOA5* and described two haplotypes that were independently associated with increased plasma triglyceride concentrations in Caucasians, African-Americans, and Hispanics [20]. Between 25 and 50% of individuals in these populations carry at least one of the two risk haplotypes, designated *APOA5*2* and *APOA5*3*.

Despite these results, questions remain about the relationship between *APOA5* and *APOC3* sequence variants and plasma triglyceride concentrations. Other than the initial results showing no association of SNP 4 (located between *APOA5* and *APOC3*) and *SstI* in *APOC3* with plasma triglyceride concentrations in a single population [1], no data are available about the linkage disequilibrium (LD) structure or haplotype patterns between *APOA5* and *APOC3*. The purpose of this study was to analyze comprehensively sequence variants across the entire apolipoprotein gene cluster and adjacent regions and to determine the LD and haplotype structure of the region to assess any potential relationship between *APOA5* and *APOC3* haplotypes and their individual alleles. The data we present support that the *APOA5* locus is separated from the other apolipoprotein genes by a region of increased recombination. Of all haplotypes across *APOA5*, only one haplotype (*APOA5*2*) is significantly associated with the minor alleles of the *APOC3 SstI* ($3238G \rightarrow C$) and the two promoter polymorphisms ($-482C \rightarrow T$ and $-455T \rightarrow C$), indicating a complex and intricate relationship between different haplotypes across this gene cluster on human chromosome 11. In contrast, the second risk haplotype (*APOA5*3*) is not associated with other haplotypes or alleles in the neighboring apolipoprotein genes, thus establishing that *APOA5* independently contributes to interindividual differences in plasma triglyceride concentrations in humans.

Results

SNP genotyping

For our analysis of linkage disequilibrium and haplotype structure, we identified 67 SNPs in the *APOA1/C3/A4/A5* region from our resequencing efforts, published reports, and public databases (dbSNP). SNPs were genotyped in individuals from 10 independent three-generation

CEPH families from Utah and included all four grandparents, both parents, and two children (one boy and one girl). After removing SNPs that were not polymorphic in our sample or were not in Hardy–Weinberg equilibrium, we analyzed 49 SNPs that spanned a total of approximately 152 kb of sequence on human chromosome 11q23, resulting in an average distance of 3112 bp between neighboring SNPs. Almost 82 kb of sequence distal to *APOA5* and over 60 kb of sequence proximal to *APOA5* were investigated. Within the *APOA5* gene (SNPs 17–23), the average distance between SNPs was 504 bp, spanning at total of 3021 bp. The flanking regions including the *APOA4/C3/A1* gene cluster had an average distance of 3620 bp between SNPs. Sixteen SNPs were located distal to *APOA5*, 26 SNPs were located proximal. Previous studies reported that LD between SNPs in different regions of the human genome extended over a range of 6–110 kb [21–23], therefore we selected a high SNP density for our study (1 SNP per 3–4 kb) to ensure our ability to detect any LD between neighboring SNPs. A complete list of all SNPs used for the analysis can be found in Table 1. The approximate location of all SNPs relative to the genes in this interval is depicted in Fig. 1.

Analysis of linkage disequilibrium across the apolipoprotein gene cluster

To determine the extent of LD in our samples, $|D'|$ was calculated for all pairs of SNPs according to Lewontin [24]. Only phased data from unrelated individuals were included in the calculations. Data included all grandparental haplotypes of the CEPH families (a total of 80 independent chromosomes), which were determined from genotyping data of the three-generation families using GENEHUNTER [25]. A schematic diagram of all pairwise comparisons between the SNPs in the region based on the CEPH samples is shown in Fig. 2. As is evident from the graphic, there is significant LD between large numbers of SNPs. Sixty-four percent of all pairwise $|D'|$ values equal 1, the maximum possible value, indicating complete LD between the two SNPs. Sixty-four percent of all pairwise $|D'|$ values are greater than 0.8. If only SNPs with a minor allele frequency greater than 10% are considered for these pairwise calculations, 18% of pairwise $|D'|$ values are 1, and 31% are greater than 0.8. Despite the extensive LD across the region, the pattern is disrupted by several neighboring SNPs with low LD (e.g., SNPs 29–32, see Fig. 2).

Haplotype structure of *APOA5*

Initially, we examined the haplotypes present in our sample across the *APOA5* gene locus. Seven SNPs (SNPs 17–23) around the gene that were included in the haplotype analysis are in complete linkage disequilibrium. These SNPs span 3021 bp and include the entire coding region of *APOA5*. The locations of all SNPs in this block and the observed haplotypes are illustrated in Fig. 3. In our CEPH sample set, only two haplotypes have a frequency over 10%, one consisting of the major allele for all seven SNPs and the other including the minor allele of SNP 18. The two previously described triglyceride-risk haplotypes (*APOA5**2 and *APOA5**3) as well as another haplotype involving the minor allele of a nonsynonymous SNP (SNP 19) are found in only 3 of the 80 chromosomes (3.8%).

Relationship of *APOA5* haplotypes to *APOC3* polymorphisms

Next, we examined the relationship between *APOA5* haplotypes and the three polymorphisms in *APOC3* (*Sst*I (3238G → C, SNP 41), –482C → T, and –455T → C (SNPs 38 and 39)). Since these SNPs have been shown to be associated with plasma triglyceride concentrations, we tried to determine whether the previously identified triglyceride risk haplotypes for *APOA5* or any other haplotypes in the *APOA5* region preferentially cosegregate with either allele of these three *APOC3* polymorphisms. The results of our analysis are presented in Table 2. The haplotype *APOA5**2 is strongly associated with the minor alleles for all three *APOC3* polymorphisms. Eighty-five percent of chromosomes with *APOA5**2 contained the minor allele for *Sst*I. In contrast, the other previously identified risk haplotype, *APOA5**3, showed

no significant association with the three minor *APOC3* alleles. Only one haplotype with a frequency of greater than 10% that spans all SNPs from *APOA5* to *APOC3* (SNPs 17–41) can be identified. In this haplotype, every SNP is represented by the common allele (data not shown). This analysis of the relationship of *APOA5* risk haplotypes and SNP alleles in and around *APOC3* suggests that the *APOA5**3 haplotype exerts its effect on triglycerides independent of *SstI* and the two promoter polymorphisms.

Analysis of recombination by the four-gamete test

If two neighboring SNPs occurred through single mutation events in human history, one would expect to see only three of four possible allele combinations in the gametes in the absence of recombination between the two SNPs. Therefore, the four-gamete test [26] has traditionally been used as a measure of evidence of recombination. However, the presence of a fourth gamete could also be caused by gene conversion or multiple recurring mutations at the site. Thus, the test has been used in other studies primarily as supporting evidence for recombination or disruption in LD between SNPs [27].

As shown in Fig. 4, the results of the four-gamete test for our CEPH sample revealed results similar to those of the LD analysis (Fig. 2). At several sites between *APOA5* and *APOC3*, a fourth gamete type can be found in this small sample, suggesting that recombination may have contributed to the breakdown in continuous LD between the two genes. As a consequence, *APOA5* (SNPs 17–23) and *APOC3* (SNPs 38–41) do not reside in the same continuous genomic region uninterrupted by recombination.

Haplotype block structure and recombination in the apolipoprotein gene cluster region

Previous studies have suggested that historic recombination events result in complex patterns of haplotype blocks across genomic regions. To support our results of the LD and four-gamete test analyses that suggested a separation of the *APOA5* and *APOC3* loci, we used a simplified approach to defining adjacent haplotype blocks. Our main emphasis was to determine whether there was any evidence for a haplotype block extending from *APOA5* to *APOC3*. Our analyses using $|D'|$ and the four-gamete test had already suggested that the two gene regions had been separated by historical recombination events. Therefore, we used a haplotype block-based analysis approach to gather additional evidence for increased recombination in the region between *APOA5* and *APOC3* compared to the remainder of the investigated genomic interval. Haplotype blocks were defined as contiguous sets of neighboring SNPs in which 90% of all chromosomes in the sample were represented by the four most common haplotypes for the given set of SNPs (block definition 1). The definition was based on the observation in previous studies that within haplotype blocks, 80–90% of all chromosomes were represented by the three or four most common haplotypes for that block [22,28–30]. For our analysis, the first block was defined around *APOA5*, and the size of the block was maximized using the block definition. Subsequently, the neighboring haplotype blocks were determined using the same block criteria, starting from the first two SNPs adjacent to the initial *APOA5* block and adding one additional SNP at a time. A representative diagram of the resulting haplotype blocks and all represented haplotypes in each block is shown in Fig. 5. A total of nine blocks covered the entire flanking region. No block included both *APOA5* and *APOC3* (SNPs 17–41). The average size of each block was 4.7 SNPs (3–7 SNPs), covering on average 13,299 bp (6343–32,111 bp).

To explore the effect of the block definition parameters on the block boundaries, we repeated the analysis on the same CEPH dataset with a different haplotype block definition in which 80% of all chromosomes were represented by the three most common haplotypes (block definition 2). With this block definition, a total of eight blocks were defined for the flanking region, containing on average 5.7 SNPs (3–9 SNPs) and covering 12,535 bp (6207–37,968 bp).

Again, no block included both *APOA5* and *APOC3*. The block boundaries based on the two different block definitions coincided only for the blocks immediately adjacent to the *APOA5* haplotype block. With increasing distance from *APOA5*, the blocks overlapped significantly for the majority of SNPs (>70%), but not entirely.

To verify that no continuous haplotype block that includes both *APOA5* and *APOC3* exists, we used the haplotype definition used by Gabriel et al. [29], which is based on measures of LD across the region. As with our analysis, no single haplotype block can be identified that spans both genes (data not shown). The same result is obtained when the HaploBlockFinder [31] is used. This approach, based on the greedy algorithm used by Patil et al. [28], also divides the entire dataset into sets of consecutive haplotype blocks. As in our approach, *APOA5* SNPs are always grouped in one block, regardless of the parameters used, and this region is separated from the *APOC3* region by several interspersed small haplotype blocks (data not shown).

Next, we attempted to identify regions where recombination was increased between haplotype blocks. If such a region separated *APOA5* from any or all of the other apolipoprotein genes, it would be highly unlikely that any association seen with an *APOA5* haplotype could be caused by a SNP or haplotype at one of the other genes.

If no recombination had occurred between adjacent haplotype blocks, it would be predicted that in every sample tested, a haplotype in one block would be paired with just one specific haplotype in the adjacent block. In reality, however, all samples having a specific haplotype in one block do usually continue with more than just one haplotype in the neighboring block. This often results in a large number of observed haplotype combinations across block boundaries. The more different haplotype combinations one would find, the higher would be the number of necessary recombination events that “mixed” the ancestral haplotypes between the two blocks.

To use this metric in our analysis, we determined the total number of “links” between all haplotypes of adjacent blocks (i.e., the total number of combinations of haplotypes from two adjacent blocks that were found in our sample set). These links are illustrated by lines in Fig. 5. We then determined the minimum number of links needed to connect each haplotype to at least one haplotype in the adjacent block (the set of links one would expect without any recombination). The difference between these two numbers represents the “excess” combinations, i.e., the number of combinations of haplotypes from adjacent blocks that arose through recombination of the ones included in the set of “minimum” combinations. We calculated the percentage of excess combinations and also the percentage of samples represented by these excess combinations between haplotype blocks.

For the CEPH haplotypes, the greatest number of samples represented by excess haplotype combinations across block boundaries could be found between SNPs 23 and 25 (22% of all samples), immediately adjacent to the *APOA5* gene. This was significantly different from the average of all other regions ($p < 0.01$). The greatest percentage of excess connections was found across the block boundary between SNP 27 and SNP 31. Almost half (49%) of the connections represented excess connections between adjacent blocks. Again, this percentage was significantly different from the average for all other block boundaries (25%, $p < 0.01$). Both statistical analyses identified the region adjacent to *APOA5* as a region of significantly increased recombination and therefore clearly separated haplotypes of *APOA5* from *APOC3* and the other apolipoprotein genes. This evidence for increased recombination between the two genes explains the lack of a contiguous haplotype block that spans both genes in our sample set.

Discussion

The two apolipoprotein genes *APOA5* and *APOC3* have repeatedly been reported to affect plasma triglyceride concentrations in humans and in mouse models. For both genes, common genetic variants that are strongly associated with alterations in human plasma lipid concentrations have been identified. Despite these results, the correlation between the two genes is not clear, and questions remain about the relationship of the SNPs in these two neighboring genes. It is possible that the association seen with SNPs in one of the two genes might be due to a causative SNP located in the other gene that is in strong LD. If this were true, one would expect to find either a single haplotype block spanning both genes or two distant haplotypes that are in strong LD. Our analysis of LD between SNPs and their resulting haplotype patterns indicates that no continuous haplotype block that would link *APOA5* and *APOC3* exists. Furthermore, we identify a region of significantly increased recombination separating the two genes, suggesting that the observed association of *APOA5* variants on plasma triglyceride concentrations is independent of *APOC3*.

Traditionally, studies investigating LD between SNPs have relied on SNPs discovered through resequencing efforts. However, recent studies have shown that public databases contain over 50% of all SNPs with a frequency >10% in the human genome [32]. Therefore, we utilized this resource in our studies and selected SNPs from public databases. Special additional emphasis was given to SNPs that were previously reported in association studies across the gene cluster region. While we did not utilize every common SNP in the region, the high degree of LD observed between adjacent SNPs makes it likely that additional SNPs located between the SNPs used in the study would also be in LD and therefore be represented by the same haplotypes. This can be illustrated by the fact that the three most common haplotypes of *APOA5* can be unambiguously identified using only publicly available SNPs and ignoring SNPs discovered by our resequencing efforts.

Our results of linkage disequilibrium from pairwise analysis of SNPs across the gene cluster region show significant allelic association across the entire cluster in the CEPH sample of individuals of Northern European descent, but the LD structure is complex, and no clear blocks are evident. While several sets of neighboring SNPs are in significant LD with each other, the results do not resolve the question whether there is strong LD between variants in *APOA5* and *APOC3*. Individual *APOA5* SNPs show strong LD to *APOC3*, but the two genes seem to be separated by a region of low LD (SNPs 29–32, see Fig. 2). The detection of high overall levels of LD (60% of pairwise $|D'| = 1$) may be partly due to low minor allele frequencies for several of the SNPs used in the analysis. Low minor allele frequencies for one or both SNPs artificially inflate $|D'|$ and increase the probability of $|D'| = 1$. Nevertheless, when only SNPs with a minor allele frequency greater than 10% are used, approximately one-third of all pairwise comparisons still detect significant LD ($|D'| > 0.8$) in the CEPH sample. This illustrates the high overall degree of LD across the entire gene cluster region. Despite the high overall LD, *APOA5* is still separated from the adjacent apolipoprotein genes by a region of low LD, suggesting increased recombination between *APOA5* and the *APOA4/C3/A1* cluster. This is supported by the results of the four-gamete test. Several pairs of adjacent SNPs between *APOA5* and the *APOA4/C3/A1* cluster show all four possible combinations of SNP alleles, suggesting recombination events mixing the ancestral alleles. While the occurrence of all four gamete types is not unambiguous proof of recombination (four gametes could also be observed in the case of recurring mutation events or gene conversion), results from the LD analysis and the four-gamete test support the hypothesis of increased recombination in the region between *APOA5* and *APOC3*.

To test this hypothesis further, as well as to analyze the haplotype block structure reflecting the complex LD patterns across the entire region, we defined haplotype blocks as areas with

limited haplotype diversity within each block. The definitions of a haplotype block in our analysis were based on previous haplotype analyses by other groups. Most studies reported that within haplotype blocks, 80–90% of the chromosomes are represented by three or four common haplotypes [22,28–30]. Accordingly, we defined haplotype blocks as contiguous sets of neighboring SNPs where 90% of all chromosomes in the sample were represented by the four most common haplotypes for the given set of SNPs (block definition 1) or 80% of the chromosomes were represented by the three most common haplotypes (block definition 2). Currently, there is extensive debate about the best approach for defining haplotype blocks in the human genome. Large-scale efforts (HapMap Project) are under way to generate comprehensive data about these blocks for the entire human genome. However, no single approach has been proposed and accepted for the analysis of these data at this point. Different studies have based their block definitions either exclusively on the degree of LD detected in that region or on a balanced approach to minimize the number of SNPs needed to differentiate all major haplotypes [33]. The resulting haplotype blocks either are contiguously subdividing the region into blocks of different size [28] or identify localized blocks of high LD [29]. Discussions have been focusing on the biological usefulness of different definitions and the usefulness of the resulting partitioning for disease association studies [34,35]. In our studies, we defined haplotype blocks exclusively for the purpose of examining and detecting possible recombination events. In addition, we tried to assess whether specific SNPs can always be found in one haplotype block regardless of the parameters used to define them.

With the two definitions used in our analysis, we find no continuous haplotype block that includes both the *APOA5* and *APOC3* gene regions. The resulting individual blocks are relatively short (approximately 13 kb), and their accurate location is different depending on the block definition used. To our knowledge, no previous study examined the effect of varying definition parameters on the resulting haplotype block structure. Our data for the apolipoprotein gene cluster region indicate that blocks obtained using different definitions overlap well for regions of high linkage disequilibrium, but the boundaries shift for regions with less LD between SNPs. In this genomic region, over 70% of SNPs are assigned to blocks that are identified with both definitions, reflecting the high degree of LD in this gene cluster. Other proposed algorithms to define the haplotype block structure [28,29] lead to the same conclusion that the *APOA5* and *APOC3* gene regions do not reside in the same haplotype block.

The blocks identified in our analysis were independent of the SNPs selected for analysis, their density, or their allele frequency. Analyzing fewer SNPs across the region, the exact location of haplotype block boundaries varied, but three genomic intervals were always included in haplotype blocks. One of these regions included *APOA5*, and the two other regions were located upstream of *APOA4* and in the 3' region of *MGC13125*. Similar results were obtained when only SNPs with a minor allele frequency above 10% were included in the analysis.

We would like to emphasize that the block structure defined by our analysis approach was used only to test whether *APOA5* and *APOC3* are located in one block. Any of a variety of other block definition algorithms yields the same result. However, we subsequently used our data for a novel approach to estimate the variation in recombination across this genomic interval to support the findings from our LD analysis and the four-gamete test. For this, we analyzed the degree of recombination found between adjacent haplotype blocks. The approach unambiguously identifies two haplotype block boundaries proximal to *APOA5* that show increased recombination. The two regions, indicated in Fig. 5 by the black arrows, clearly separate *APOA5* from the neighboring apolipoprotein genes. Previous studies have already estimated that the apolipoprotein gene cluster region has a fourfold higher average recombination rate than the genome-wide average [36]. However, due to the limited resolution of RFLP analysis, the exact location for this increased recombination was not identified. It is conceivable that the region between *APOA5* and *APOA4* is at least in part responsible for this

increased overall recombination rate. This increased recombination supports that the association found between SNPs and haplotypes in *APOA5* and altered plasma triglyceride concentrations is not due to linkage disequilibrium with causative *APOC3* variants.

Despite the LD and haplotype evidence that separates *APOA5* from the other apolipoprotein genes in the cluster, individual *APOA5* haplotypes do show association with *APOC3* alleles. We identified five different haplotypes for the *APOA5* region in the CEPH sample. Of these, only risk haplotype *APOA5**2 is significantly associated with the minor allele of the *SstI* polymorphism and the two promoter variants in *APOC3*. All other haplotypes show little or no association with the minor allele, and one haplotype is exclusively found with the common allele (S1) of *SstI*. This association is also evident in other populations that have been ascertained for these polymorphisms (data not shown). In another Caucasian population from the Midwest United States [37], the frequencies of the two risk haplotypes are higher than in our CEPH sample (4.6 and 8.0% for *APOA5**2 and *APOA5**3, respectively). Still, *APOA5**2 is clearly associated with the minor allele of *SstI* (72.7% of all chromosomes with the *APOA5**2 risk haplotype).

Overall, this analysis confirms that the other previously identified risk haplotype (*APOA5**3) shows no association with *SstI* or the *APOC3* promoter SNPs, suggesting that the association with plasma triglycerides is likely to be independent of *APOC3* or other genes in this cluster, although we could not test directly for this in our sample.

It remains to be seen whether the effect of the *APOA5**2 risk haplotype is due to a functional polymorphism in *APOC3* or whether the *APOA5* haplotype is responsible for the association seen between the minor allele of the three *APOC3* SNPs and plasma triglyceride concentrations. In addition, the relationship between the two genes should be investigated in greater detail in other ethnic groups as well, given the apparent difference in the frequency of *APOA5* risk haplotypes [20]. Functional studies are under way to address this question. Our analysis clearly shows that *APOA5* is an independent risk gene for elevated plasma triglyceride concentrations as evidenced by the *APOA5**3 risk haplotype. Extensive functional studies will be needed to elucidate the individual contributions of SNPs within *APOA5* to altered plasma triglyceride concentrations.

Materials and methods

Genotyping of single nucleotide polymorphisms

In total, 67 SNPs were selected for our analysis from a region spanning 152.5 kb around the apolipoprotein gene cluster on human chromosome 11. Of these SNPs, 7 in and around *APOA5* were initially discovered by us and other collaborators by direct sequencing [1,20]. An additional 4 SNPs in the region were discovered through resequencing, and the remaining 56 SNPs were retrieved from dbSNP and published reports. Currently, 348 SNPs are listed in dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/index.html>) for the apolipoprotein gene cluster region. We selected all SNPs in the *APOA5* region, and 1 SNP per approximately 3 kb of flanking sequence to ensure a dense map to detect LD between neighboring SNPs. In addition, we included SNPs in and around the other apolipoprotein genes that were reported previously in association studies: SNPs 34, 35, and 36 are nonsynonymous changes in *APOA4*; SNPs 38 (*APOC3* -482), 39 (*APOC3* -455), and 41 (*SstI*) are located in *APOC3*; and SNP 43 (*XmnI* polymorphism) is located in *APOA1* (see Table 1).

All SNPs were genotyped using the biplex Invader assay, as previously described [38]. SNPs were genotyped on a set of 80 samples representing 10 independent three-generation CEPH families from Utah. Of the 67 SNPs, 9 SNPs (13.4%) were not polymorphic in the samples used for our study, and three assays failed. An additional 3 SNPs (4.5%) had minor allele

frequencies of less than 1% in our samples, and were not included in the analysis. For all SNPs, we tested for Hardy–Weinberg equilibrium. Three SNPs significantly deviated from equilibrium and were excluded from further analysis. The remaining 49 SNPs (73.1%) were used for all subsequent analyses and are numbered in Table 1.

Analysis of linkage disequilibrium

Genotypes obtained from the CEPH families were imported into GENEHUNTER [25]. The HAPLO program option was used to determine the grandparental haplotypes across the entire region. Less than 0.4% genotyping errors were detected, and all erroneous genotypes were repeated and unambiguously resolved. Haplotype data for all independent samples (grandparents) were used to calculate pairwise linkage disequilibrium. Lewontin's parameter of LD, $|D'|$ [24], was calculated for all pairwise comparisons.

Four-gamete test

The four-gamete test was applied to the CEPH dataset as described in previous studies [26, 27]. For any given two-SNP haplotype AB , mutation will lead to the formation of either Ab or aB . A haplotype consisting of the two alleles ab can arise only through recombination or repeated mutation. The four-gamete test essentially examines the sample set for the presence of all or a reduced number of gametic allele combinations. The test was applied as implemented in the software program DnaSP 3.51 [39].

Analysis of the haplotype block structure

To define haplotype blocks, the region in and around *APOA5* was investigated first. SNPs 17–23 were defined as the starting block for our algorithm using MATLAB. Haplotype blocks were defined as regions where at least 90% of all haplotypes were represented by the four most common haplotypes. Alternatively, the definitions were changed so that 80% of all haplotypes were represented by the three most common haplotypes. In the first iteration, the initial block of SNPs 17–23 was expanded by adding 1 neighboring SNP at a time from either the 3' or the 5' end to maximize the size of the region. Once the maximal block size around *APOA5* was determined, the algorithm was started again to determine blocks in the neighboring SNPs. Here, the algorithm started with the first 2 SNPs outside the initial *APOA5* block and then added the third SNP and tested whether the haplotype block definition was still fulfilled. If the four major 3-SNP haplotypes still represented more than 90% of all chromosomes in the sample, the next SNP was added. The iterations stopped once the haplotype definition criteria were no longer met. The last set of SNPs fulfilling the conditions was defined as a new haplotype block, and the algorithm was restarted using the first 2 SNPs outside the newly defined block. Haplotype blocks were defined using the complete set of 49 SNPs, a set of SNPs comprising all SNPs in the initial *APOA5* haplotype block and all SNPs in the flanking regions with minor allele frequencies above 10%, and finally a set of SNPs consisting of the *APOA5* SNPs and only every second SNP in the flanking regions. These two modified datasets were used to determine the effects of allele frequencies and SNP density on the definition of haplotype blocks and their boundaries in our sample.

Analysis of recombination

Based on the haplotype block structure determined as described above, we counted all combinations of haplotypes across each boundary that could be found in our sample. Each combination is represented by a line in Fig. 4. We counted the total number of lines for each boundary. Then we determined the minimum number of combinations needed to connect each haplotype in one block with at least one haplotype in the other block. These lines represent the minimum number of combinations needed to explain the observed data. With only these combinations of haplotypes, the sample set would still have all haplotypes for both blocks

represented. Any additional connections found in the sample would have arisen through recombination between any of these “essential” haplotype combinations. Next we proceeded to count the additional (excess) combinations of haplotypes found in our sample and the number of chromosomes with these combinations. Calculating the ratio of these additional lines to the total number of lines (i.e., percentage excess connections between haplotype blocks) and determining the percentage of samples that are represented by these excess connections give a measure of recombination between the two blocks. Due to the non-normality of the data, the significance of difference in the dataset was assessed using nonparametric tests (Wilcoxon, Median, Kolmogorov–Smirnov).

Acknowledgments

We thank D. Savic and L. Smith (Medical College of Wisconsin) for excellent technical assistance in genotyping, and T. Wang (Medical College of Wisconsin) for statistical support. Thanks to S. Schaffner (Whitehead Institute) for graciously analyzing our data using the LD-based haplotype block algorithm. This work was supported in part by the Biological and Environmental Research Program, the U.S. Department of Energy’s Office of Science, NIH Grant HL0748168 (M.O.), the NIH–NHLBI Programs for Genomic Application Grant HL66681 (E.M.R.), and NIH Grant HL071954A (E.M.R., L.A.P.) through the U.S. Department of Energy under Contract DE-AC03-76SF00098.

References

1. Pennacchio LA, et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 2001;294(5540):169–173. [PubMed: 11588264]
2. Plump AS, et al. ApoA-I knockout mice: characterization of HDL metabolism in homozygotes and identification of a post-RNA mechanism of apoA-I up-regulation in heterozygotes. *J Lipid Res* 1997;38(5):1033–1047. [PubMed: 9186920]
3. Maeda N, et al. Targeted disruption of the apolipoprotein C-III gene in mice results in hypotriglyceridemia and protection from postprandial hypertriglyceridemia. *J Biol Chem* 1994;269(38):23610–23616. [PubMed: 8089130]
4. Groenendijk M, et al. The apoA1-CIII-AIV gene cluster. *Atherosclerosis* 2001;157(10):1–11. [PubMed: 11427198]
5. Dammerman M, et al. An apolipoprotein CIII haplotype protective against hypertriglyceridemia is specified by promoter and 3’ untranslated region polymorphisms. *Proc Natl Acad Sci USA* 1993;90(10):4562–4566. [PubMed: 8099442]
6. Zeng Q, et al. An apolipoprotein CIII marker associated with hypertriglyceridemia in Caucasians also confers increased risk in a west Japanese population. *Hum Genet* 1995;95(4):371–375. [PubMed: 7705829]
7. Shoulders CC, et al. Variation at the apo A1/CIII/AIV gene complex is associated with elevated plasma levels of apo CIII. *Atherosclerosis* 1991;87(2–3):239–247. [PubMed: 1906714]
8. Shoulders CC, et al. Characterization of genetic markers in the 5’ flanking region of the apo A1 gene. *Hum Genet* 1993;91(2):197–198. [PubMed: 8096496]
9. Ordovas JM, et al. Restriction fragment length polymorphisms of the apolipoprotein A-I, C-III, A-IV gene locus. Relationships with lipids, apolipoproteins, and premature coronary artery disease. *Atherosclerosis* 1991;87(1):75–86. [PubMed: 1678604]
10. Surguchov AP, et al. Polymorphic markers in apolipoprotein C-III gene flanking regions and hypertriglyceridemia. *Arterioscler Thromb Vasc Biol* 1996;16(8):941–947. [PubMed: 8696957]
11. Stocks J, Paul H, Galton D. Haplotypes identified by DNA restriction-fragment-length polymorphisms in the A-I C-III A-IV gene region and hypertriglyceridemia. *Am J Hum Genet* 1987;41(2):106–118. [PubMed: 2887109]
12. Tybjaerg-Hansen A, et al. Genetic markers in the apo A1-CIII-AIV gene cluster for combined hyperlipidemia and preposition to atherosclerosis. *Atherosclerosis* 1993;100(2):157–169. [PubMed: 8102848]
13. Paul H, Galton D, Stocks J. DNA polymorphic patterns and haplotype arrangements of the apo A-1, apo C-III, apo A-IV gene cluster in different ethnic groups. *Hum Genet* 1987;75(3):264–268. [PubMed: 2881879]

14. Tas S. Strong association of a single nucleotide substitution in the 3'-untranslated region of the apolipoprotein-CIII gene with common hypertriglyceridemia in Arabs. *Clin Chem* 1989;35(2):256–259. [PubMed: 2914370]
15. Ahn YI, et al. DNA polymorphisms of the apolipoprotein AI/CIII/AIV gene cluster influence plasma cholesterol and triglyceride levels in the Mayans of the Yucatan Peninsula, Mexico. *Hum Hered* 1991;41(5):281–299. [PubMed: 1778603]
16. Endo K, et al. Association found between the promoter region polymorphism in the apolipoprotein A –V gene and the serum triglyceride level in Japanese schoolchildren. *Hum Genet* 2002;111(6):570–572. [PubMed: 12436249]
17. Nabika T, et al. The genetic effect of the apoprotein AV gene on the serum triglyceride level in Japanese. *Atherosclerosis* 2002;165(2):201–204. [PubMed: 12417270]
18. Ribalta J, et al. Newly identified apolipoprotein AV gene predisposes to high plasma triglycerides in familial combined hyperlipidemia. *Clin Chem* 2002;48(9):1597–1600. [PubMed: 12194944]
19. Talmud PJ, et al. Relative contribution of variation within the *APOC3/A4/A5* gene cluster in determining plasma triglycerides. *Hum Mol Genet* 2002;11(24):3039–3046. [PubMed: 12417525]
20. Pennacchio LA, et al. Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. *Hum Mol Genet* 2002;11(24):3031–3038. [PubMed: 12417524]
21. Reich DE, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411(6834):199–204. [PubMed: 11346797]
22. Olivier M, et al. Complex high-resolution linkage disequilibrium and haplotype patterns of single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21. *Genomics* 2001;78(1–2):64–72. [PubMed: 11707074]
23. Abecasis GR, et al. Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001;68(1):191–197. [PubMed: 11083947]
24. Lewontin RC. On measures of gametic disequilibrium. *Genetics* 1988;120(3):849–852. [PubMed: 3224810]
25. Kruglyak L, et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58(6):1347–1363. [PubMed: 8651312]
26. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 1985;111(1):147–164. [PubMed: 4029609]
27. Bonnen PE, et al. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res* 2002;12(12):1846–1853. [PubMed: 12466288]
28. Patil N, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;294(5547):1719–1723. [PubMed: 11721056]
29. Gabriel SB, et al. The structure of haplotype blocks in the human genome. *Science* 2002;296(5576):2225–2229. [PubMed: 12029063]
30. Daly MJ, et al. High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29(2):229–232. [PubMed: 11586305]
31. Zhang K, Jin L. HaploBlockFinder: haplotype block analyses. *Bio-informatics* 2003;19(10):1300–1301.
32. Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. *Nat Genet* 2003;33(4):457–458. [PubMed: 12652301]
33. Olivier M. A haplotype map of the human genome. *Physiol Genom* 2003;13(1):3–9.
34. Wall JD, Pritchard JK. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet* 2003;73(3):502–515. [PubMed: 12916017]
35. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003;4(8):587–597. [PubMed: 12897771]
36. Antonarakis SE, et al. DNA polymorphism haplotypes of the human apolipoprotein *APOA1-APOC3-APOA4* gene cluster. *Hum Genet* 1988;80(3):265–273. [PubMed: 2903847]
37. Kissebah AH, et al. Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. *Proc Natl Acad Sci USA* 2000;97(26):14478–14483. [PubMed: 11121050]
38. Olivier M, et al. High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Res* 2002;30(12):e53. [PubMed: 12060691]

39. Rozas J, Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bio-informatics* 1999;15(2):174–175.

Fig. 1.

Diagram of the apolipoprotein gene cluster region on human chromosome 11q23. All genes in this region are depicted by horizontal block arrows. Apolipoprotein genes are highlighted in white and all gene names are given below the arrows. The approximate location of each SNP is indicated by the vertical arrows. Numbers correspond to the SNP numbering in Table 1.

Fig. 2.

Pairwise linkage disequilibrium. Pairwise LD was calculated according to Lewontin [24]. $|D'|$, the normalized LD measure, is depicted for all pairwise comparisons. $|D'|$ values of 1 are indicated by black squares, white squares indicate $|D'|$ values of less than 0.5. The approximate location of SNPs is shown by the lines connecting individual rows and columns to the diagram of the genomic region.

Fig. 3.

APOA5 haplotypes. The locations of the seven SNPs comprising the *APOA5* haplotype block are illustrated on the top. Gene exons are indicated by boxes, black regions indicate the coding region of the gene. Haplotypes are depicted by the diagrams in the lower half. White squares indicate the major allele, black squares the minor allele of a SNP. The names of three haplotypes to the left of the diagram refer to the nomenclature used by Pennacchio et al. [20].

Fig. 4.

Four-gamete test. For all pairwise comparisons, pairs of SNPs with all four gametic allele combinations represented in the CEPH sample are depicted by black squares, pairs of SNPs with three gametes are illustrated by white squares. The black arrow indicates an area of two pairs of neighboring SNPs between *APOA5* and *APOC3* where all four gametes are present. This suggests a possible area of recombination between the two gene regions.

Fig. 5.

Haplotype patterns in the CEPH sample. The top illustrates the gene location across the region. Below, all 49 SNPs are represented by consecutive squares. White squares indicate the major allele, while black or gray squares highlight the minor allele of an individual SNP. Haplotype blocks were identified using block definition 1. Lines between haplotype blocks represent connections found in the CEPH samples. The blocks are numbered consecutively, and approximate locations of the block boundaries are illustrated by the connecting lines. The location of the *APOA5* haplotype block (block 5) is shaded, and the locations of the three *APOC3* SNPs are illustrated by the arrows in blocks 8 and 9. When we examined the CEPH chromosomes for all 49 SNPs together, we identified 33 chromosomes that shared the complete 49-SNP haplotype with at least one other chromosome. The block patterns of these chromosomes are highlighted in dark in the haplotype blocks (black squares). All other haplotypes in each block are shown in gray.

Table 1

SNPs used for analysis

Position	dbSNP	Note	MAF (%)	Other name	Source
1	rs1240783		48.8		dbSNP
2	rs180368		19.2		dbSNP
3	rs180363		32.5		dbSNP
4	rs1813805		12.5		http://pga.lbl.gov/SNP/APOA1C3A4A5.html
5	rs18120998		2.5	ApoA1/A5.34529	http://pga.lbl.gov/SNP/APOA1C3A4A5.html
	rs180338	No SNP			dbSNP
	rs586248	No SNP			dbSNP
6	rs180329		47.5		dbSNP
7	rs180324		20.0		dbSNP
8	rs2075295		23.8		dbSNP
9	rs918143		47.5		dbSNP
10	rs918144		43.8		dbSNP
11	rs2187126		8.8		dbSNP
12	rs1268353		28.8		dbSNP
13	rs664059		37.5		dbSNP
14	rs2041967		14.7		dbSNP
15	rs1043740	No SNP			dbSNP
16	rs1942478		7.5		dbSNP
17	rs603446		48.8		http://pga.lbl.gov/SNP/APOA1C3A4A5.html
18	rs2266788		19.0	ApoA5-SNP1	dbSNP
	rs619054	Rare			dbSNP
19	rs2075291				dbSNP
	rs3135507		3.8	ApoA5-V153M	dbSNP
20	rs2072560		3.8	ApoA5-SNP2	http://pga.lbl.gov/SNP/APOA1C3A4A5.html
21	rs3135506		3.8	ApoA5-S19W	dbSNP
22	rs651821		3.8	ApoA5-Kozak	http://pga.lbl.gov/SNP/APOA1C3A4A5.html
	rs648450	Rare			dbSNP
23	rs662799		3.8	ApoA5-SNP3	http://pga.lbl.gov/SNP/APOA1C3A4A5.html
24	rs1787680		46.3		dbSNP
25	rs1729410		38.8		dbSNP
26	rs633389		2.5		dbSNP
27	rs633867		2.5		dbSNP
	rs672143	No SNP			dbSNP
	rs682109	HWE			dbSNP
	rs1263163	No SNP			dbSNP
28	rs625524		2.5		dbSNP
29	rs1729408		36.3		http://pga.lbl.gov/SNP/APOA1C3A4A5.html
30	rs1263166		20.0		dbSNP
31	rs1263171		21.8		dbSNP
	rs2727793	No SNP			dbSNP
32	rs2849168		20.0		dbSNP
33	rs2849165		33.8		dbSNP
34	rs1263177		6.3	ApoA4-Q360H	dbSNP
35	rs5110		20.0	ApoA4-T347S	dbSNP
36	rs675		4.1	ApoA4-A141S	dbSNP
	rs5104				dbSNP
	rs5091	Rare	27.5		dbSNP
37	rs2098453				dbSNP
	rs2542052	Fail			dbSNP
38	rs2854117		20.3	ApoC3(-482)	dbSNP
39	rs2854116		35.0	ApoC3(-455)	dbSNP
40	rs5132		1.3		dbSNP
41	rs5128		6.3	ApoC3-SsrI	dbSNP

	Position	dbSNP	Note	MAF (%)Other name	Source
	118218480	rs5081	Fail		dbSNP
	118218911	rs5078	No SNP		dbSNP
	118219070	rs4882	No SNP		dbSNP
	118219086	rs5077	No SNP		dbSNP
	118220657	rs2727786	Fail		dbSNP
42	118223275	rs2727784		ApoA1- <i>Pst</i> I 27.5	dbSNP
43	118223484			5.0ApoA1- <i>Xmn</i> I	Genbank Acc. X67732.1 GI:28768
44	118225966	rs494606		9.0	dbSNP
45	118229482	rs614944		1.3	dbSNP
46	118231910	rs2289893		16.3	dbSNP
47	118234992	rs583219		8.8	dbSNP
48	118236025	rs598503		2.5	http://pga.lbl.gov/SNP/APOA1C3A4A5.html
	118236053	rs888246	HWE		dbSNP
49	118237124	rs640411		1.3	dbSNP

Listed are the number of the SNP when used in the LD and haplotype analysis (first column), the position of each SNP based on the Golden Path genome annotation (human assembly November 2002), the dbSNP reference number, the minor allele frequency (MAF), and other names that were used to describe the SNP, as well as the source of the SNP information.

Table 2

Association of *APOA5* haplotypes with *APOC3* polymorphisms

<i>APOA5</i> haplotype	Haplotype frequency	<i>APOC3 SstI</i>		<i>APOC3 –482</i>		<i>APOC3 –455</i>	
		1	2	1	2	1	2
1111111 (<i>APOA5*1</i>)	69.2%	53	1	48	4	44	10
1211111	19.2%	15	0	9	6	2	13
2112122 (<i>APOA5*2</i>)	3.8%	0	3	0	3	0	3
1111211 (<i>APOA5*3</i>)	3.8%	3	0	3	0	3	0
1121111	3.8%	2	1	1	2	1	2

Listed are the *APOA5* haplotypes found in our populations. “1” indicates the major allele of the SNP, “2” indicates the minor allele of the SNP. Haplotypes are listed in order from SNP 17 through 23. Haplotype frequencies are given for both samples, and the number of chromosomes carrying each of the alleles of the three *APOC3* polymorphisms is listed for each *APOA5* haplotype.