Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels

David A. Hinds,[1] Albert B. Seymour,[2] L. Kathryn Durham,[3] Poulabi Banerjee,[2] Dennis G. Ballinger,[1]
Patrice M. Milos,[2] David R. Cox,[1] John F. Thompson[2] and Kelly A. Frazer[1]*

[1] Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 94043, USA
[2] Genomic and Proteomic Sciences, Pfizer Global Research and Development, Eastern Point Road, Groton, CT 06340, USA
[3] Biostatistical Applications, Pfizer Global Research and Development, Eastern Point Road, Groton, CT 06340, USA
* *Correspondence to*: Tel: +1 (0) 650 625 4504; Fax: +1 (0) 650 625 4510; E-mail: kelly_frazer@perlegen.com

Date received (in revised form): 10th August 2004

Abstract

Association studies are used to identify genetic determinants of complex human traits of medical interest. With the large number of validated single nucleotide polymorphisms (SNPs) currently available, two limiting factors in association studies are genotyping capability and costs. Pooled DNA genotyping has been proposed as an efficient means of screening SNPs for allele frequency differences in case-control studies and for prioritising them for subsequent individual genotyping analysis. Here, we apply quantitative pooled genotyping followed by individual genotyping and replication to identify associations with human serum high-density lipoprotein (HDL) cholesterol levels. The DNA from individuals with low and high HDL cholesterol levels was pooled separately, each pool was amplified by polymerase chain reaction in triplicate and each amplified product was separately hybridised to a high-density oligonucleotide array. Allele frequency differences between case and control groups with low and high HDL cholesterol levels were estimated for 7,283 SNPs distributed across 71 candidate gene regions spanning a total of 17.1 megabases. A novel method was developed to take advantage of independently derived haplotype map information to improve the pooled estimates of allele frequency differences. A subset of SNPs with the largest estimated allele frequency differences between low and high HDL cholesterol groups was chosen for individual genotyping in the study population, as well as in a separate replication population. Four SNPs in a single haplotype block within the cholesteryl ester transfer protein (*CETP*) gene interval were significantly associated with HDL cholesterol levels in both populations. Our study is among the first to demonstrate the application of pooled genotyping followed by confirmation with individual genotyping to identify genetic determinants of a complex trait.

Keywords: *association study, HDL cholesterol, CETP, SNPs, pooled genotyping, haplotypes*

Introduction

Association studies are widely viewed as one of the most promising methods for identifying the genetic determinants of human phenotypic traits of medical interest, such as common diseases and individual responses to the drugs used to treat those diseases.[1] Therefore, a considerable amount of research has been focused on developing methodologies that efficiently screen candidate gene regions or whole genomes for associations of complex phenotypes with genetic markers, such as single nucleotide polymorphisms (SNPs). The methodology relies on having a set of common genetic markers at a sufficiently dense coverage across the genome, such that either the causal variant itself or a marker in linkage disequilibrium with the causal variant will be tested in the association study. Thus, to be comprehensive and reproducible, a whole genome scan study requires the assay of hundreds of thousands of densely spaced SNP markers in a large number of samples. There is a considerable body of experimental[2–6] and theoretical[7–9] work that suggests genotyping of pools consisting of DNA from many individuals is a viable alternative to individual genotyping. Pooled assays replace many measurements of individual samples with a few measurements of a pooled sample — with a corresponding reduction in cost, time and labour. Here, we describe one of the first large-scale SNP association studies in which this methodology has been applied and validated.

Human population studies have shown that serum high density lipoprotein (HDL) cholesterol concentrations are inversely correlated with the development of premature coronary heart disease.[10] In this report, we describe a two-stage study to identify genetic markers associated with HDL cholesterol levels. First, we use a pooled genotyping screen to identify SNPs likely to have large frequency differences between low and high HDL cholesterol groups. Starting from 7,283 SNPs distributed across 71 candidate regions, we use the pooled data to select about 300 SNPs with the strongest evidence for association. We then individually genotype these SNPs, to confirm their allele frequency differences in the low and high HDL cholesterol individuals in the study group. We confirm associations identified in the study population by individually genotyping SNPs with significant allele frequency differences in a replicate population.

We also describe a novel method for using independently derived haplotype map data to improve the power of an association study based on pooled genotyping. Using genotype data from a separate set of ethnically diverse individuals, we determine haplotype blocks and sets of common haplotype patterns that together account for most of the variation in a given genomic interval. From pooled genotype data, we estimate frequency differences for these common patterns between case and control groups. These pattern differences enable us to make more accurate estimates of the individual SNP frequency differences that exploit redundancies in the haplotype map, thereby reducing experimental error in the individual SNP measurements.

Materials and methods

SNP discovery and haplotype map construction

In an independent, previously described study, a genome-wide SNP collection was obtained using high-density oligonucleotide array-based resequencing.[11] Briefly, we generated somatic cell hybrids by fusing lymphoblast cell lines from the Coriell Polymorphism Discovery Resource[12] with a hamster cell line to form between 20 and 50 haploid somatic cell hybrids for each human chromosome. DNA was isolated and amplified by long-range polymerase chain reaction (PCR), and the PCR products were fragmented, labelled and hybridised to a series of SNP discovery arrays. These arrays were designed such that each base of the reference sequence was queried by eight 25-mer probes. We identified SNPs from the resulting fluorescence intensity data using a pattern recognition algorithm.

We used a dynamic programming algorithm[13] to partition these haploid SNP discovery data into haplotype blocks. SNPs having minor allele frequencies of at least 10 per cent in the SNP discovery data were included in the map. We required all blocks to satisfy the condition that at least 80 per cent of the haploid samples could be assigned to common haplotype patterns having greater than 10 per cent frequency. For a block having N common haplotype patterns, we also required at least $N - 1$ patterns to have tagging SNPs that distinguished each of those patterns from all of the others.

Sample selection

The study population was derived from a cohort of individuals (self-reported Caucasian) from the ACCESS study,[14] which was made up of males, postmenopausal females and premenopausal females that either had, or were at risk for, cardiovascular disease. Whole blood from subjects participating in this study was obtained in accordance with the Declaration of Helsinki (2000) of the World Medical Association, in addition to appropriate informed consent documentation defining the study design and providing an assessment of the risks and benefits associated with study participation. Individuals with high and low HDL cholesterol levels were selected as the top and bottom 15 per cent of the continuous HDL distribution from each group, resulting in the following samples: 166 high HDL ($\geq$54.9 mg/dl) and 182 low HDL ($\leq$36.1 mg/dl) males; 140 high HDL ($\geq$64.0 mg/dl) and 142 low HDL ($\leq$47.3 mg/dl) postmenopausal females; and 17 high HDL ($\geq$67.4 mg/dl) and 24 low HDL ($\leq$42.2 mg/dl) premenopausal females. HDL cholesterol was measured in fasting samples from four preclinical visits, all DNA samples were collected at baseline (ie without drug treatment). In this population, the interaction between age and HDL did not warrant an adjustment for age in the selection of cases and controls.

The replicate population consisted of 83 low HDL and 78 high HDL samples from postmenopausal women (self-reported Caucasians). These samples represented the 25 per cent tails from both ends of the continuous HDL distribution of an independent cohort from an osteoporosis study (the cohort was not selected on the basis of their HDL cholesterol levels or other cardiovascular risk factors), with the high HDL cut-off at 62 mg/dl, the low HDL cut-off at 42 mg/dl and a mean age of 54.4 years.

Construction of DNA pools

We constructed four DNA pools for estimation of SNP allele frequency differences between the low and high HDL cholesterol groups. Five of the 671 samples of the study population were excluded from pooled genotyping due to insufficient amount of DNA or failed normalisation. After removal of these samples, there were 345 low HDL samples and 321 high HDL samples remaining. The low HDL cholesterol samples were randomly split into two subgroups and used to construct pool A (consisting of 173 individuals) and pool B (consisting of 172 individuals). Likewise, the high HDL cholesterol samples were randomly split into two subgroups and used to construct pool C (consisting of 161 individuals) and pool D (consisting of 160 individuals).

Genomic DNA was extracted from whole blood using the PureGene DNA isolation system (Gentra) per manufacturer's protocol. DNA samples were quantified using a PicoGreen assay kit (Molecular Probes) and SpectraFluor Plus Tecan plate reader according to the manufacturers' instructions, and then

Table 1. Seventy candidate genes analysed in the association study.

ABCB6	ABCC3	ABHD1	ACACA	ACACB	ACAT1	ACAT2
ADAM28	ADAMTS4	AR	ATP6V1E1	B3GALT2	BPI	CACNG5
CEL	CLN2	CRP	CTSC	EPHB6	ESR2	ESRRB
FPR1	G6PC	GHSR	GJA4	GLRA1	GPRC5B	HAT
HDL-CBP	HNF4A	KCNJ9	LIPE	LIPG	LOC51275	MBTPS1
MBTPS2	MMP3	MTP	NDRG1	NR0B2	NR1H2	NR1H3
NR1H4	NR1I3	NR2F1	PCOLCE2	PLA2G2A	PLTP	PON1
PON2	PPARA	PPARD	PTPRF	PTPRH	RORA	RORB
RPS6KC1	SAA1	SAA2	SAA4	SCAP	SCD	SERPINE1
SPTLC1	SPTLC2	SSTR1	TCF1	TRHR	TSHR	UPS3

diluted to a standard concentration using a Packard Multi-Probe Robot. Equimolar aliquots of DNA were transferred into one of four pool tubes using a Packard MultiProbe robot. Each pool was then requantified by PicoGreen assay and the pools diluted to 20 ng/μl for use as a PCR template.

SNP selection for pooled genotyping

We selected 71 gene targets based on a variety of criteria. One gene, the cholesteryl ester transfer protein (*CETP*), had been previously shown to be associated with HDL cholesterol[10,15,16] and served as a positive control. The remaining 70 candidate genes (Table 1; see also supplementary Table 1; supplementary tables have been posted at: www.perlegen.com/newsroom/supplemental/human_genomics/10_04/index.htm) were either known or suspected to be involved in lipid metabolism. We did not include some genes previously shown to be associated with HDL cholesterol levels because our goal was to identify novel associations; for example, we did not include hepatic triglyceride lipase (*LIPC*), lipoprotein lipase (*LPL*), low-density lipoprotein cholesterol receptor (*LDLR*) or ATP-binding cassette transporter A1 (*ABCA1*).[17] For the 70 candidate genes, we selected SNPs within the genomic DNA sequences encoding the transcripts, as well as 80 kilobases (kb) upstream and downstream of each transcript. We examined a larger region, spanning 1.5 megabases (Mb) upstream and downstream of *CETP*. The targeted 17.1 Mb of DNA sequence included 50 partial and 180 complete transcripts in addition to the 71 selected candidates, based on the National Center for Biotechnology Information Build 30 (see supplementary Table 2). We identified 7,283 SNP markers in these regions, at an average density of one SNP every 2.3 kb. Of these, 112 were in transcribed sequences of the 71 candidate genes, 180 were in the transcribed sequences of the 230 non-candidate genes in the intervals examined and 72 represented amino acid changes (supplementary Table 2). More than 50 per cent of the 17.1 Mb is covered by inter-SNP intervals of 10 kb or less and more than 80 per cent is in inter-SNP intervals of less than 50 kb. The 71 selected intervals contain 955 haplotype blocks, having an average of about six common SNPs and three common haplotype patterns per block.

High-density oligonucleotide arrays

High-density oligonucleotide arrays were designed so that each SNP would be interrogated by 80 25-mer oligonucleotide probes synthesised on a glass substrate. These 80 features consisted of four sets of 20 features, corresponding to reference and alternate alleles for forward and reverse strands. A set of 20 features consisted of five sets of four probes, with offsets of -2, -1, 0, $+1$ and $+2$ bases between the centre of the 25-mer probe and the SNP position. For each offset, we tiled features for each of four nucleotides substituted for the centre position of the 25-mer probe, thus at each offset we had one perfect match feature and three mismatch probes for the corresponding SNP allele (Figure 1).

Determination of pooled allele frequency estimates

For pooled genotyping, 7.25 ng genomic DNA (pooled samples) was amplified using long-range PCR reactions, pooled, labelled, hybridised to high-density arrays, stained and detected as described.[11] The four DNA pools (low HDL pools A and B and high HDL pools C and D) were each amplified by PCR using 1,222 long-range primer pairs in three replicates. The 12 sets of PCR products were hybridised to separate chips.

The fluorescence intensities of the reference and alternate perfect-match features on an array were correlated with the concentration of the corresponding SNP allele in the DNA sample. Our estimates of allele frequency, $\hat{p}$, were computed from ratios of trimmed means of intensities of the

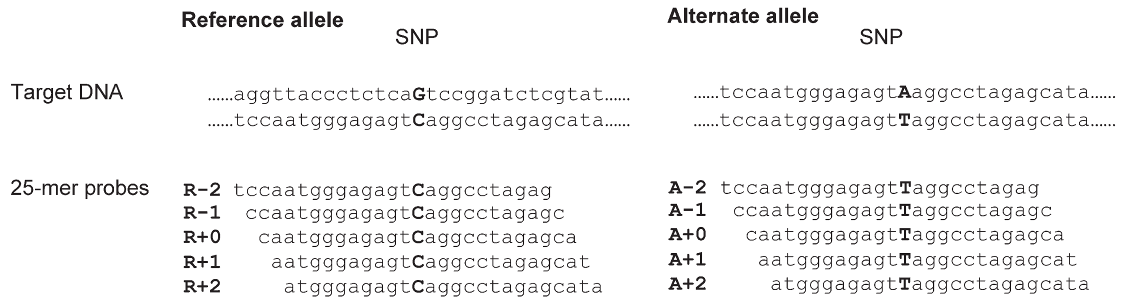
Reference allele
SNP
Alternate allele
SNP
Target DNA
......aggttaccctctcaGtccggatctcgtat......
......tccaatgggagagtCaggcctagagcata......
......tccaatgggagagtAaggcctagagcata......
......tccaatgggagagtTaggcctagagcata......
25-mer probes
R-2 tccaatgggagagtCaggcctagag
R-1 ccaatgggagagtCaggcctagagc
R+0 caatgggagagtCaggcctagagca
R+1 aatgggagagtCaggcctagagcat
R+2 atgggagagtCaggcctagagcata
A-2 tccaatgggagagtTaggcctagag
A-1 ccaatgggagagtTaggcctagagc
A+0 caatgggagagtTaggcctagagca
A+1 aatgggagagtTaggcctagagcat
A+2 atgggagagtTaggcctagagcata

Figure 1. Genotyping single nucleotide polymorphisms (SNPs) using high-density oligonucleotide arrays. Each SNP is queried by 80 25-mer oligonucleotides synthesised on a glass substrate. The ten oligonucleotides shown are perfect-match probes for the reference (R) and alternate (A) alleles at five offsets on the forward strand sequence relative to the SNP (−2, −1, 0, +1, +2). Not shown are additional mismatch probes where the middle positions of the probes shown are replaced by the three alternate nucleotides, and an equivalent set of probes for the reverse strand.

perfect-match features after subtracting a measure of background computed from trimmed means of intensities of mismatch features

$$\hat{p} = \frac{\tilde{I}_{PM,Ref} - \tilde{I}_{MM}}{(\tilde{I}_{PM,Ref} - \tilde{I}_{MM}) + (\tilde{I}_{PM,Alt} - \tilde{I}_{MM})}$$

where:

$$\tilde{I}_{MM} = \tfrac{1}{4}(\tilde{I}_{MM,Ref,Fwd} + \tilde{I}_{MM,Ref,Rev} + \tilde{I}_{MM,Alt,Fwd} + \tilde{I}_{MM,Alt,Rev})$$

$$\tilde{I}_{PM,Ref} = \tfrac{1}{2}(\tilde{I}_{PM,Ref,Fwd} + \tilde{I}_{PM,Ref,Rev})$$

$$\tilde{I}_{PM,Alt} = \tfrac{1}{2}(\tilde{I}_{PM,Alt,Fwd} + \tilde{I}_{PM,Alt,Rev})$$

The $\tilde{I}$ terms denote trimmed mean intensities for a set of features denoted by the subscript. The trimmed means are arithmetic means of the intensity measurements after discarding the highest and lowest 25 per cent of values. In cases where this did not yield an integer number of terms, one more term was included and the smallest and the largest terms received half weight. Each set of 20 features contributed five perfect-match measurements for one allele, one perfect-match measurement for the other allele at offset 0, and 14 mismatch measurements. Thus, for example, there were six perfect-match features for the reference allele on the forward strand, and:

$$\tilde{I}_{PM,Ref,Fwd} = \tfrac{1}{3}\big(\tfrac{1}{2} I_{PM,Ref,Fwd,2} + I_{PM,Ref,Fwd,3} + I_{PM,Ref,Fwd,4} + \tfrac{1}{2} I_{PM,Ref,Fwd,5}\big)$$

where the numeric subscripts denote positions in the list of six sorted intensities.

Two quality control metrics were used to assess the reliability of the intensities for a SNP in an array scan. The first metric, 'conformance', measured the presence of specific target DNA for that SNP. The second metric, signal to background ratio, measured the relative amounts of specific and non-specific binding. Cut-offs were applied to both metrics, and SNP feature sets that did not pass either metric were discarded from further analysis.

Conformance was computed independently for both reference and alternate allele feature sets, and a maximum taken of the two values. The conformance for a particular allele was defined as the fraction of feature sets for which the perfect-match feature was brighter than all three mismatch features. In the 80-feature SNP tiling, each allele had ten such sets of four features. SNP measurements having conformance <0.9 were discarded from further evaluations.

The signal to background ratio was calculated from intensity measurements for both alleles, for the perfect-match versus mismatch features, as:

$$signal = \sqrt{\tilde{I}^2_{PM,Ref} + \tilde{I}^2_{PM,Alt}}$$

$$background = \sqrt{\tilde{I}^2_{MM,Ref} + \tilde{I}^2_{MM,Alt}}$$

The trimmed mean intensities for perfect-match and mismatch feature sets were obtained as described above. SNP measurements having signal/background <1.5 were discarded from further evaluations.

For each SNP, we obtained a total of 12 allele frequency estimates, $\hat{p}$, as three independent measurements for each of the four DNA pools. Estimated allele frequency differences, $\Delta\hat{p}$, between low and high HDL groups were determined from averages of the replicates for each pool:

$$\Delta\hat{p} = \tfrac{1}{2}(\hat{p}_A + \hat{p}_B) - \tfrac{1}{2}(\hat{p}_C + \hat{p}_D)$$

Haplotype block fitting algorithm

In order to limit the number of SNPs requiring subsequent genotyping in individual samples, we developed an analysis method that used our independently derived haplotype

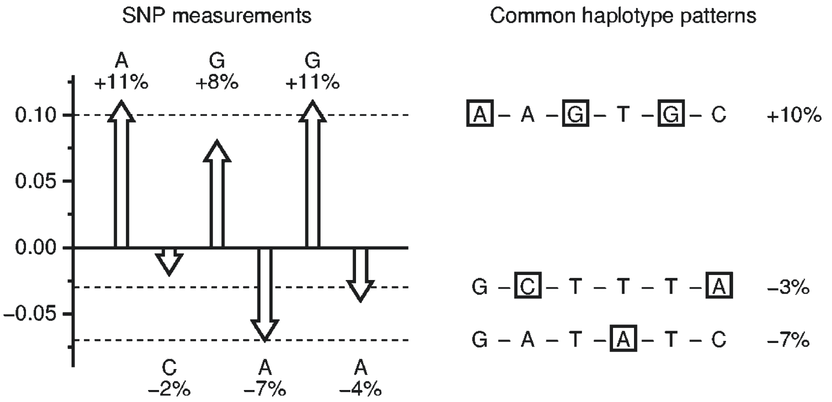

Figure 2. Haplotype block-fitting analysis. Starting from estimated allele frequency differences for each individual single nucleotide polymorphism (SNP) from pooled genotyping, we use linear regression to solve for frequency differences of the underlying common haplotype patterns. In this example, we show a hypothetical haplotype block consisting of six SNPs and three common haplotypes. Measured frequency differences are shown for the haplotype tagging alleles for each SNP, which are also indicated by boxes in the haplotype patterns. In this example, we are estimating two free parameters from six SNP measurements, since the three pattern differences are constrained to sum to 0. Thus, these pattern differences should have lower variance than the individual SNP measurements. From the pattern differences, we are able to estimate the true allele frequency differences for each SNP more accurately.

map information to refine estimates of SNP allele frequency differences between pooled DNA samples in case-control studies. The method exploits the fact that within a haplotype block, most of the variation in SNP allele frequencies can be accounted for by variation in the frequencies of a relatively small set of common haplotype patterns — defined as patterns present at a frequency of at least 10 per cent in the ethnically diverse population used for SNP discovery. Within a block, the sum of differences in these pattern frequencies between two groups should be approximately 0, to the extent that those patterns in the haplotype map accurately represent the total genetic diversity of that interval (Figure 2).

The method uses linear regression to determine these underlying haplotype pattern frequency differences, given a set of estimated SNP allele frequency differences for a haplotype block. Our method for haplotype map construction guarantees that in every block, there are at least enough SNPs to determine the frequencies of the common haplotype patterns. Most SNPs are in blocks that contain additional redundant SNPs, so if measurement errors are uncorrelated, regression should yield estimates that are more accurate than the original SNP measurements. From the fitted pattern differences, more accurate estimates of the true allele frequency differences for individual SNPs can then be determined.

Let $\Delta\hat{p}_i$ be the estimated frequency difference of the 'reference' alleles for SNP i within a haplotype block, and let Δf_j be the (unknown) frequency difference of common haplotype pattern $j \in 1\ldots N$. Our model proposes that:

$$\Delta\hat{p}_i = \sum_{j=1}^{N} m_{ij}\Delta f_j + \varepsilon$$

where m_{ij} is a coefficient that takes a value of $+0.5$ if the allele at position i in pattern j matches the reference allele and -0.5 if it matches the alternate allele for that SNP. The reason for the 0.5 factor is that the frequency difference for an allele would otherwise be double counted when differences for the complete set of patterns are evaluated. We further require that the pattern frequency differences must sum to 0; this constraint can be folded into the previous equation by eliminating Δf_N and defining $r_{ij} \equiv m_{ij} - m_{iN}$ to obtain:

$$\Delta\hat{p}_i = \sum_{j=1}^{N-1} r_{ij}\Delta f_j + \varepsilon$$

Solving these equations given $\Delta\hat{p}_i$ and r_{ij} is a linear regression problem. Standard regression statistics (R^2 and the P value for an F test) can be used to judge the quality of the fit of the SNP data to the haplotype pattern information. Deviations from a perfect fit can arise both from experimental errors and inaccuracies in the haplotype model. In instances where the quality of the fit to the haplotype map is good, the fitted allele frequency differences should have lower variance than the raw data for individual SNPs because they incorporate information about the expected correlations between SNPs.

A similar method could be used to estimate haplotype pattern frequencies in each pooled sample, with a constraint that the frequencies of common patterns add up to 1. We chose to work in the space of allele frequency differences for several reasons. The frequency differences are the quantities we are ultimately interested in, and it seemed most parsimonious to evaluate a fit for these differences directly, rather than performing separate fits on frequencies in each pool and then combining these to obtain differences. Also, the quality of a fit on absolute frequencies would be sensitive to the presence of rare haplotypes not included in the model, even under the null hypothesis of no pool differences. Our constraint on frequency differences summing to 0 only implies that the proportion of rare haplotypes in case and control pools is similar. Finally, due to experimental differences in SNP hybridisation characteristics, we have more confidence in our ability to detect pool differences than to obtain unbiased estimates of absolute allele frequencies.

Determination of individual genotypes

For individual genotyping by high-density oligonucleotide arrays, samples were amplified by short-range multiplex PCR, labelled, hybridised to the arrays, stained and detected as described.[18]

The individual genotypes for an SNP were determined by clustering measurements from multiple scans in the two-dimensional space defined by reference and alternate perfect-match trimmed mean intensities. Trimmed mean intensities were computed as described above. We used a K-means algorithm to assign $\hat{p}$ measurements to clusters representing distinct diploid genotypes. Instead of estimating the background intensity term $\tilde{I}_{MM}$ from a single scan, we determined an optimal value for each SNP that minimised the variance in $\hat{p}$ within the assigned genotype clusters. The K-means and background optimisation steps were iterated until cluster membership and background estimates converged. To determine the appropriate number of genotype clusters, we repeated the analysis for one, two and three clusters and selected the most likely solution, considering likelihoods of the data and the cluster parameters. The data likelihood was determined using a normal mixture model for the distribution of $\hat{p}$ around the cluster means. The model likelihood was calculated using a prior distribution of expected cluster positions (ie homozygous reference allele near $\hat{p} = 1.0$, heterozygote near $\hat{p} = 0.5$ and homozygous alternate allele near $\hat{p} = 0.0$).

For individual genotyping by template-directed dye-terminator incorporation with fluorescence-polarisation detection (FP-TDI),[19] samples were amplified by PCR, primer extension was performed using AcycloPrime FP SNP detection kit (Perkin Elmer Life Sciences) and changes in fluorescence polarisation were measured using Analyst HT (LJL Biosystems) as described.[16]

Results

Population stratification analysis

In an association study, systematic differences in ancestry between case and control groups can produce large numbers of statistically significant but spurious associations.[20,21] We examined the 348 individuals with low HDL levels and the 329 individuals with high HDL levels in the study population to ensure that they were adequately matched prior to constructing DNA pools. We individually genotyped the samples for 300 SNPs that are genetically unlinked and uniformly spaced across the genome, as described previously.[18]

In χ^2 tests for association with the HDL cholesterol phenotype, we observed a small excess of moderate p values. For 280 SNPs giving high-quality genotype data, 43 had $p < 0.1$ versus 28 expected. A sensitive global test for population structure based on the sum of χ^2 statistics[22] was significant ($p < 0.001$); however, a permutation analysis of the genotype data indicated that the expected increase in variance of allele frequency measurements due to stratification of this magnitude was less than 1 per cent. We also analysed the genotype data for population structure using the *structure* program.[23] The *structure* program uses a model-based clustering method for identifying subpopulations such that, within a cluster, all markers are in Hardy–Weinberg and linkage equilibrium. This analysis did not show convincing evidence for more than one subpopulation. In runs with between two and five assumed clusters, most samples were assigned similar admixture proportions in each predicted subpopulation; for two clusters, 75 per cent of samples were given admixture proportions between 0.4 and 0.6. Based on these results, and the limited accuracy of pooled genotyping assays, we judged that the low and high HDL cholesterol groups were adequately matched.

Pooled genotyping results

For each SNP, we estimated an allele frequency difference, $\Delta\hat{p}$, between the low HDL cholesterol and high HDL cholesterol pools. We then excluded a small proportion of the pooled data due to spurious experimental errors, such as saturated features or inconsistent hybridisation patterns. We also excluded SNPs where all three measurements for any one of the four pools failed and SNPs where the standard error of $\Delta\hat{p}$ exceeded 0.025. Of the 7,283 SNPs tiled on the array, 6,611 (91 per cent) passed all of these data quality filters.

Haplotype block fitting analysis

Of the 6,611 SNPs for which we obtained good pooled genotyping data, 4,387 SNPs were included in the haplotype map. Table 2 shows the results of the haplotype block fitting analysis for these SNPs; the results for all blocks, the subset of blocks that are informative (those that contain redundant SNP information) and the subset of these that had $p < 0.05$ in an F test for agreement of the $\Delta\hat{p}$ with the haplotype model for that block are shown. Good fits should only be possible for blocks that have real allele frequency differences between the low and high HDL cholesterol pools, either due to sampling variation or association with the phenotype. Thus, we would expect most blocks to have poor p values, due to a lack of significant allele frequency differences. In fact, more than 40 per cent of the 4,387 SNPs are in blocks with good agreement between

Table 2. Haplotype block-fitting results and analysis of variance.

	All blocks	Informative[a]	$p < 0.05$[b]
Haplotype blocks	955	507	172
SNPs passing quality filters	4387	3758	1934
SNPs contributing to fits	4171	3578	1833
Fitted degrees of freedom	1736	1143	442
Residual degrees of freedom	2435	2435	1391
% degrees of freedom used	42%	32%	24%
Total sum of squares	2.241	1.947	1.258
Fitted sum of squares	1.725	1.431	1.002
Residual sum of squares	0.516	0.516	0.256
% variance explained	77%	73%	80%

[a] Blocks having redundant information, ie at least as many SNP measurements as common haplotype patterns.
[b] Informative blocks for which an *F* test on the fit of the SNP data to the haplotype structure had $p < 0.05$.
SNPs, single nucleotide polymorphisms.

$\Delta\hat{p}$ and the haplotype model, and these tend to be the larger blocks. Uninformative blocks often contain just one or two SNPs and while they represent a large fraction of all blocks, they represent a much smaller proportion of SNPs and base pairs covered. Here, informative blocks represented 53 per cent of all blocks, but included 86 per cent of SNPs in the haplotype map and about 75 per cent of the DNA sequence.

Analysis of variance allows us to determine how much of the variation in SNP allele frequencies observed between the DNA pools is consistent with the haplotype map and how much is residual variation due to experimental errors in the $\Delta\hat{p}$ measurements, the contribution of rare patterns not represented in the haplotype map and errors in the haplotype map. We can measure the effectiveness of the algorithm by the extent to which the fraction of variance explained by the fitted haplotype patterns exceeds the fraction of degrees of freedom used in the fits. In this analysis (Table 2), we found that about 77 per cent of all the variance in the data was consistent with the model based on common haplotypes. Based on the number of free parameters in the haplotype model, we would have expected only 42 per cent of the variance to be accounted for by chance. We repeated this analysis after permuting the individual $\Delta\hat{p}$ measurements. Here, the haplotype map explained only 43 per cent of the variance and only 5 per cent of SNPs were in blocks having $p < 0.05$ in an *F* test. This analysis shows that the agreement between the haplotype model and the original $\Delta\hat{p}$ data could not arise by chance.

Selection of SNPs for individual genotyping

Selecting the SNP markers that are the most likely to have large allele frequency differences based on the pooled array data is difficult. The set of SNPs having the largest absolute $\Delta\hat{p}$ is dominated by a subset of measurements with very high experimental variance. A *t*-test is also inadequate, because it favours SNPs with low experimental variance, even if the $\Delta\hat{p}$ is too small to be of biological interest and is probably due to sampling variation. The experimental variance is poorly determined from the limited number of data points available. Due to differences in SNP calibration in our genotyping assay, our ability to estimate absolute allele frequencies, and hence sampling variance, is similarly limited. Based on data from experiments with pools of known composition, we found that the strategy of excluding data for SNPs with very high standard errors, and then selecting SNPs with the largest $\Delta\hat{p}$, performed as well or better than tests based on variance estimates for each SNP (data not shown).

A total of 312 SNP markers were chosen for individual genotyping based on the capacity of a small high-density oligonucleotide array. Based on the pooled allele frequency data, we selected 284 SNPs expected to have large allele frequency differences. Half of the 284 SNPs were chosen to be ‘haplotype conforming’ — belonging to informative haplotype blocks that had good fits with $p < 0.05$ — while the other half were chosen to be ‘non-conforming’ SNPs selected from the remainder based only on pooled estimates of $\Delta\hat{p}$. We ranked 1,934 conforming SNPs by the smaller of their actual and fitted $\Delta\hat{p}$ values, and selected the top 142 SNPs yielding a cut-off of $|\Delta\hat{p}| > 0.037$. For 4,677 non-conforming SNPs, ranking by absolute $\Delta\hat{p}$ and selecting the top 142 yielded a cut-off of $|\Delta\hat{p}| > 0.048$. We selected a higher proportion of conforming SNPs for individual genotyping because their consistency with the haplotype map provided additional evidence for allele frequency differences at those positions. We did include non-conforming SNPs, however, so as not to overlook signals that were not in blocks, or for which the fit to the haplotype map was poor for other reasons. An additional 28 SNPs that did not meet these criteria were selected because they were either in candidate loci of interest or had been independently genotyped in the same population using fluorescence polarisation. They were used to assess the accuracy of our high-density array-based individual genotyping.

Individual genotyping data quality analysis

A total of 832 DNA samples in the study and replicate populations were individually genotyped for the 312 selected SNPS. Three quality-control filters were applied to the individual genotype data. We first required that SNPs have an unambiguous genotype call in at least 80 per cent of the 832 DNA samples assayed. Secondly, we required that both SNP alleles be segregating in the population (ie have at least two identifiable genotype clusters). Finally, we required that the SNP alleles be in Hardy–Weinberg equilibrium ($p > 0.001$). We found that large deviations from Hardy–Weinberg equilibrium were generally associated with systematic hybridisation artefacts. Of the 312 SNPs, 284 (91 per cent) passed all three data quality filters.

To estimate the quality of the individual genotypes called using the high-density oligonucleotide array platform, we compared our genotype calls with those obtained using FP-TDI for 19 SNPs in three gene regions (*CETP*, endothelial lipase [*LIPG*] and liver receptor alpha [*LXRα*]). The call rate (the fraction of assigned genotypes out of potential genotypes) for the array platform is above 98 per cent, very similar to that generated using FP-TDI using the same DNA samples (supplementary Table 3). Of the genotypes called by both methods, the concordance (the fraction of SNPs assigned genotypes by both methods that were in agreement) between the oligonucleotide array and FP-TDI methodologies is greater than 99 per cent.

Evaluation of the pooled genotyping screen

To evaluate the effectiveness of the pooled genotyping step for estimating allele frequency differences between the case and control DNA pools, we examined the relationship between pooled allele frequency estimates and allele frequencies determined by individual genotyping. For each of the 284 SNPs selected from the pooled data, we have allele frequency estimates for four pooled samples and corresponding individual genotype data for all the samples used to compose the pools. While the numbers of data points and ranges of allele frequencies for each SNP are small, we can still use these data to examine the relationship between a pooled $\hat{p}$ and an allele frequency p determined by individual genotyping. This relationship for an individual SNP is very nearly linear; however, there is substantial variation in slope and intercept between SNPs (Figure 3). A regression of the $\hat{p}$, averaged over the four HDL pools against an allele frequency p, determined

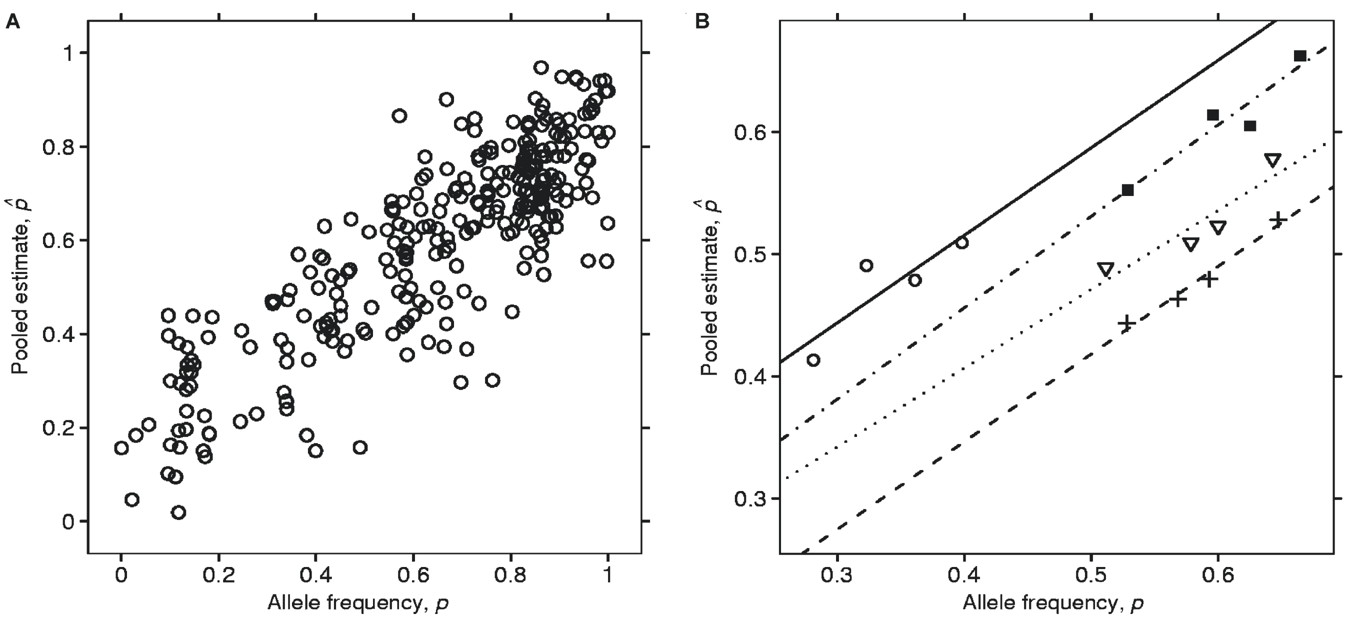

Figure 3. Relationship between pooled allele frequency estimates and allele frequencies determined by individual genotyping. The frequency estimates from pooled genotyping, $\hat{p}$, are linearly related to allele frequencies, p, determined by individual genotyping. (**A**) Across all SNPs that were individually genotyped, variation in slope and intercept partially obscures the strength of this relationship. Here, we show $\hat{p}$ plotted against p averaged over the four high-density lipoprotein pools for 284 SNPs. (**B**) For each SNP, we have independent measurements of p and $\hat{p}$ in four pools. We show representative data for four SNPs having (due to sampling variation or association) relatively large separation between the four values of p.

Table 3. Distribution of χ^2 test scores in SNPs selected for individual genotyping.

	All SNPs	Conforming[a]	Non-conforming[b]
Selected SNPs	284	142	142
Number meeting quality criteria	263	135	128
$p < 0.04$[c]	90	59	31
$p < 0.01$[c]	41	24	17

[a] SNPs selected based on corroborating evidence from the haplotype-fitting procedure.
[b] SNPs selected based on a large $\Delta\hat{p}$ but without supporting haplotype information.
[c] SNPs meeting this threshold in a χ^2 test of allelic association with the HDL phenotype.
SNPs, single nucleotide polymorphisms.

by individual genotyping for all 284 SNPs, has an R^2 of 0.71. When we examined the independent measurements of p and $\hat{p}$ in the four pools for the 284 individual SNPs, the median $R^2 = 0.85$. In principle, we could calibrate assays for each SNP using samples of known allele frequency; however, this becomes impractical when many thousands of assays are analysed. Some variation in sensitivity can be tolerated because the pooled data are only used as a screen for selecting SNP candidates for individual genotyping.

To evaluate the sensitivity of SNP selection from pooled genotyping, we used χ^2 tests to measure allelic association of each SNP with the HDL cholesterol phenotype. To the extent that pooled allele frequency differences are predictive, we should see an excess of small p values in tests for the 284 SNPs selected based on the pooled results. Given tests of N SNPs at a threshold of $p < X$, we expect $(N \times X)$ SNPs to meet that threshold due to sampling variation in allele frequencies between the pools. In fact, we see far more small p values than would be expected based on 284 tests (Table 3). From the entire 6,611 SNPs we used to choose the 284 for individual genotyping, we would expect $6{,}611 \times 0.01 = 66$ to have $p < 0.01$, and $6{,}611 \times 0.04 = 284$ to have $p < 0.04$. A perfect SNP selection procedure would have captured all of these. In fact, we captured 32 per cent of the expected total number of SNPs at the $p < 0.04$ level, and 62 per cent of the expected number at the $p < 0.01$ level. Thus, the pooled assay has sufficient sensitivity to capture a substantial fraction of SNPs having even very modest allele frequency differences at the level of sampling variation. Sensitivity for larger allele frequency differences indicative of causal associations should be correspondingly higher.

To assess the effectiveness of the haplotype fitting procedure, we looked at the numbers of 'haplotype conforming' and 'non-conforming' SNPs having small χ^2 test p values in individual genotyping (Table 3). Of SNPs having $p < 0.04$, about 65 per cent came from the 'haplotype conforming' category, and this excess was quite significant ($p \approx 0.001$). The same trend was seen among SNPs having $p < 0.01$; however, the numbers of observations were insufficient to reach statistical significance. Thus, SNPs selected based on corroborating haplotype data indeed seem to be more likely to show larger allele frequency differences in individual genotyping.

To further examine the impact of using haplotype block information to improve estimates of SNP allele frequency differences between pooled DNA samples in case-control studies, we compared the correspondence of estimates of $\Delta\hat{p}$ to actual allele frequency differences determined by individual genotyping (Figure 4). While the correspondence of pooled estimates of $\Delta\hat{p}$ to actual allele frequency differences was good ($r^2 = 0.82$), the correspondence of the haplotype-fitted estimates of $\Delta\hat{p}$ to actual allele frequency differences was substantially better ($r^2 = 0.90$). These results demonstrate that the accuracy of estimating SNP allele frequency differences in pooled genotyping is improved using haplotype block information.

SNP associations with HDL levels in the study population

Our two-stage experimental design posed a tricky multiple testing problem. While we performed tests on just 312 individually genotyped SNPs, these were selected as likely to have large allele frequency differences from a total of 6,611 SNPs with good pooled genotyping data quality. If our pooled assay was perfect, then we were effectively testing all 6,611 SNPs; if the pooled estimates were uncorrelated with allele frequencies, then we are really only testing the 312 SNPs selected for individual genotyping. Based on our capture rates for SNPs with small p values, we consider that the effective number of tests we were performing was a substantial fraction of 6,611.

Using a conservative Bonferroni correction, a global false-positive rate of 0.05 for 6,611 SNPs tested would require $p < 7.6 \times 10^{-6}$ for an association to be significant. Considering only the study cohort used for pooled genotyping, there were six SNPs, all in the *CETP* gene, that met this threshold of significance (Table 4). Of these six SNPs, four (*rs711752*, *rs708272*, *rs11508026*, *rs7205804*) had been selected based on positive results in the pooled genotyping screen and two (*rs1800775*, *rs11076175*) had been included to test cross-technology concordance of genotype calling. The four SNPs that had been assayed in the pooled screen were all in the same haplotype block (Figure 5), which had a fitted p value of ~ 0.002 in our haplotype analysis of the pooled data, and the largest fitted $\Delta\hat{p}$ values observed in the study. The

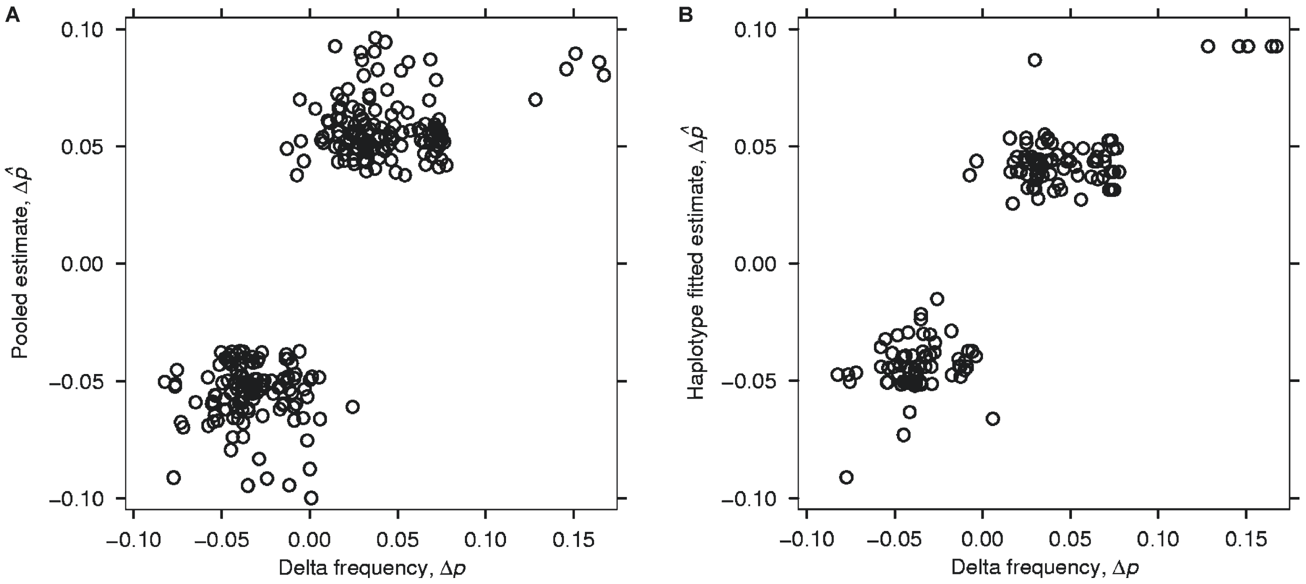

Figure 4. Correspondence between estimated pooled allele frequency differences and actual differences determined by individual genotyping. For this analysis, we included data for single nucleotide polymorphisms (SNPs) having call rates of at least 90 per cent. (A) The correlation between estimated and actual allele frequency differences for 267 SNPs is 0.82. (B) For 164 SNPs in haplotype blocks with fitted $p < 0.05$, the correlation between allele frequency difference and the estimate based on the haplotype-fitting procedure is 0.90. In the upper right corner of each plot are five SNPs in the genomic interval encoding *CETP*.

next best SNP in the data for the study cohort outside the *CETP* gene had a χ^2 test p value of 0.0015, which would not be significant, even using a generous threshold based on 312 independent tests.

SNP associations in the replicate population

To replicate an association in the study cohort, we only need to consider tests in the replicate samples for the subset of SNPs that gave significant χ^2 scores in that analysis. In the replicate population, there are only five SNPs with good data quality having $p < 0.01$, and all are in *CETP* (Table 5). For the six SNPs in *CETP* that showed significance association in the study samples, a conservative Bonferroni correction for a global false-positive level of 0.05 would require $p < 8.3 \times 10^{-3}$ to count as a replication. Of these six SNPs, four (*rs711752*, *rs708272*, *rs11508026*, *rs7205804*) were also associated at this significance level in the replicate population. Two SNPs (*rs1800775*, *rs11076175*) that had small p values in the study population did not meet significance levels in the replicate DNA samples. These differences were not unexpected, given the limited sample size of the replicate cohort and incomplete linkage disequilibrium of the SNPs in this interval.

Table 4. SNPs having significant association with HDL cholesterol in the study population.

Chromosome	Position[a]	dbSNP	ref_L[b]	alt_L	ref_H	alt_H	Δp[c]	χ^2	p value
16	56729856	rs1800775	392	196	288	322	0.193	46.2	1.1×10^{-11}
16	56730908	rs708272	446	234	318	320	0.151	33.5	7.2×10^{-9}
16	56730831	rs711752	451	239	314	312	0.165	31.2	2.4×10^{-8}
16	56733948	rs11508026	441	227	303	289	0.167	28.6	9.1×10^{-8}
16	56739509	rs7205804	436	244	312	318	0.146	28.4	9.7×10^{-8}
16	56740998	rs11076175	535	155	563	73	−0.101	28.1	1.2×10^{-7}

[a] Position on GenBank sequence *NC_000016.4*.
[b] Counts of reference and alternate alleles for low and high HDL cholesterol groups.
[c] Allele frequency difference, calculated as $\text{ref}_L/(\text{ref}_L + \text{alt}_L) - \text{ref}_H/(\text{ref}_H + \text{alt}_H)$.
dbSNPs single nucleotide polymorphism database; HDL, high-density lipoprotein.

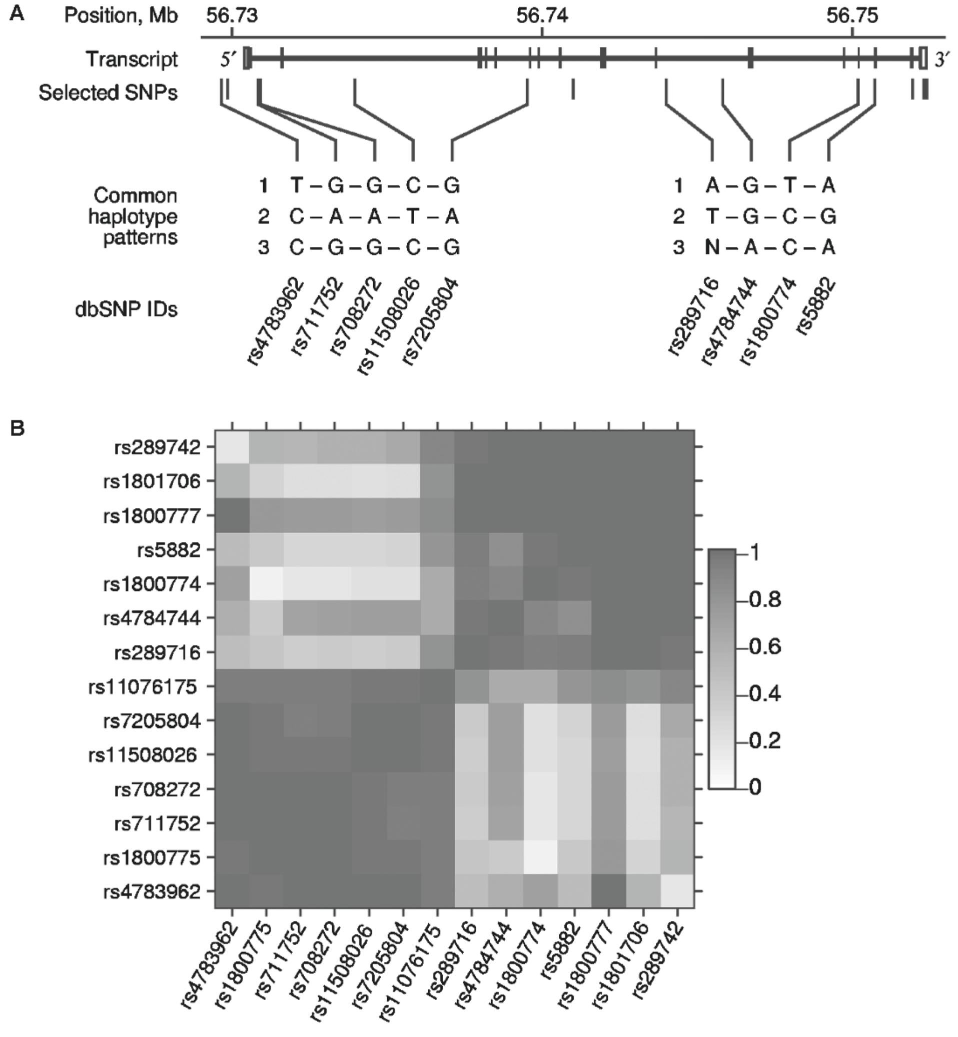

Figure 5. Single nucleotide polymorphisms (SNPs) and haplotype structure in the Cholesteryl ester transfer protein (*CETP*) gene. (**A**) The CETP transcript (GenBank accession *NM_000078*) is roughly 22 kilobases and contains 16 exons. We obtained good quality genotype data for 14 SNPs in this region. Of these, nine were common SNPs that were included in our haplotype map. In the map, these were assigned to two haplotype blocks, each having three common haplotype patterns. The 5′ block includes the Taq1B SNP, *rs708272*.[28,29] (**B**) We measured linkage disequilibrium, *D*′, for all pairs of SNPs in this interval, using an expectation maximisation (EM) algorithm to determine haplotypes. These data also show two strong blocks of high disequilibrium.

Linkage disequilibrium across the *CETP* locus

Previous studies have identified two major blocks of linkage disequilibrium across the *CETP* locus.[4,17,24] Of the 14 SNP markers in *CETP* that we examined for association with HDL cholesterol levels (supplementary Table 4), nine are members of our whole-genome haplotype map (Figure 5). Consistent with these other studies, in our map, these nine SNPs are divided into two haplotype blocks, each having three common haplotype patterns. We computed *D*′ for all pairs of the 14 SNPs in *CETP*, using an expectation maximisation (EM) algorithm to determine haplotype frequencies.[25,26] These results, again, show two blocks of very strong disequilibrium.

Discussion

The goal of this study was to determine the effectiveness of a large-scale pooled genotyping screen to identify common variants associated with a complex trait. *CETP*, which transfers cholesteryl esters from the anti-atherogenic HDL to the pro-atherogenic very-low- and low-density lipoprotein fractions, plays an important role in HDL cholesterol metabolism and served as our positive control. Correlations between SNPs in the genomic interval encoding *CETP* and variations in the mass/activity of the CETP protein and corresponding HDL levels have been intensively studied[10,16,27] and consistently

Table 5. Genotyping results for 14 *CETP* gene SNPs in the test and replicate samples.

Position[a]	dbSNP	Study samples			Replicate samples			Overall *p* value[b]
		Δp	*p* value	Odds ratio[c]	Δp	*p* value	Odds ratio	
56729658	rs4783962	0.060	1.2×10^{-2}	1.38	0.100	2.8×10^{-2}	1.87	1.4×10^{-3}
56729856	rs1800775	0.193	1.1×10^{-11}	2.26	0.104	1.5×10^{-1}	1.40	1.8×10^{-11}
56730831	**rs711752**	**0.165**	$\mathbf{2.4 \times 10^{-8}}$	**1.88**	**0.208**	$\mathbf{5.0 \times 10^{-4}}$	**2.21**	$\mathbf{5.9 \times 10^{-11}}$
56730908	**rs708272**	**0.151**	$\mathbf{7.2 \times 10^{-9}}$	**1.92**	**0.202**	$\mathbf{3.2 \times 10^{-4}}$	**2.25**	$\mathbf{1.2 \times 10^{-11}}$
56733948	**rs11508026**	**0.167**	$\mathbf{9.0 \times 10^{-8}}$	**1.85**	**0.186**	$\mathbf{1.9 \times 10^{-3}}$	**2.06**	$\mathbf{7.1 \times 10^{-10}}$
56739509	**rs7205804**	**0.146**	$\mathbf{9.7 \times 10^{-8}}$	**1.82**	**0.142**	$\mathbf{5.6 \times 10^{-3}}$	**1.89**	$\mathbf{1.9 \times 10^{-9}}$
56740998	rs11076175	−0.101	1.2×10^{-7}	0.45	−0.006	8.9×10^{-1}	0.96	1.6×10^{-6}
56743996	rs289716	−0.060	1.6×10^{-2}	0.75	−0.039	3.9×10^{-1}	0.81	1.1×10^{-2}
56745805	rs4784744	−0.075	5.0×10^{-3}	0.72	−0.157	8.7×10^{-3}	0.53	2.4×10^{-4}
56750165	rs1800774	0.009	7.5×10^{-1}	1.04	0.102	5.8×10^{-2}	1.53	2.5×10^{-1}
56750712	rs5882	−0.073	4.4×10^{-3}	0.72	−0.017	8.0×10^{-1}	0.94	7.7×10^{-3}
56751939	rs1800777	−0.041	1.1×10^{-3}	0.33	−0.009	2.9×10^{-1}	1.06	4.2×10^{-3}
56752282	rs1801706	0.029	2.0×10^{-1}	1.20	−0.022	2.9×10^{-1}	0.73	5.0×10^{-1}
56752382	rs289742	−0.023	1.2×10^{-1}	0.76	−0.053	4.6×10^{-2}	0.40	3.0×10^{-1}

[a] Position on GenBank sequence *NC_000016.4*.

[b] Calculated from a χ^2 test for the combined study and replicate samples.

[c] Calculated as $(\mathrm{ref}_L/\mathrm{alt}_L)/(\mathrm{ref}_H/\mathrm{alt}_H)$, where these are allele counts as in Table 4.

SNPs in bold are those with significant association with HDL cholesterol in both the study and replicate populations.

CETP, cholesteryl ester transfer protein dbSNP, single nucleotide polymorphism database; HDL, high-density lipoprotein.

shown to be associated. *CETP* is estimated to account for ~5 per cent of the variability of HDL levels in the general population.[17] Here, four SNPs in *CETP* had strong signals and were independently identified as being associated with HDL levels in the pooled screen. The fact that we identified *CETP* in this study as being associated with HDL levels confirms that pooled genotyping can be used in genetic association studies to identify genes underlying complex phenotypes.

While we find *CETP* to be replicable and convincingly associated with HDL cholesterol serum levels, none of the 70 candidate genes or 230 other genes in the 17.1 Mb of DNA screened appear to play a major role in the genetic variability of HDL cholesterol levels in this population. Based on the strong association of *CETP* with HDL observed in our study, we are likely to have had sufficient power to identify similar effect sizes in the other candidate genes. Recent work suggests that there are likely to be several additional genes that contribute to HDL phenotypic variance and are as yet unidentified.[17] We examined SNPs distributed across only about 0.5 per cent of the genome, and thus it is likely that these unidentified genes are located in genomic intervals that we did not examine.

In our candidate region study, we used a design incorporating stratification analysis, pooled genotyping, confirmation of promising candidate loci by individual genotyping and replication in an independent cohort. We have demonstrated that independently derived haplotype map information can be used to improve SNP selection from a pooled genotyping screen. High-density oligonucleotide arrays permit the scale and efficiency required for very large-scale association studies. These experimental methods and analysis strategies can be directly scaled up to whole-genome association studies.

Acknowledgments

We would like to thank Charles Shear and the Lipitor Team for assistance in the clinical trials; Pascual Starink, Ellen Jacobs and Eric Buljan for LIMs development and support; Geoff Nilsen, Michael Jen and Wade Barrett for designing the high-density arrays and assistance with data analysis; John Sheehan for quality analysis of the high-density array data; Albert Yee, Reed George, Julie Marschner, Joe Karbowski and Mike Kennemer for the DNA normalisation and pooling; Patrick Chu for the pooled genotyping; Clariza LaRosa, Matt Morenzoni, Pei-Hua Wang, Rajal Patel, Rhode Vergara, Robin Li, Thai Lai, Vincent Mendoza, Karel Konvicka and Renee Stokowski for the high-density array individual genotyping; Maruja Lira, Amy Mank-Seymour and Jodi Richmond for the fluorescence polarisation genotyping; Erica Beilharz for assistance with manuscript preparation; Paul Feeney and Michael Swietek for assistance in sample handling; and the patients for donating samples for research.

References

1. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
2. Germer, S., Holland, M.J. and Higuchi, R. (2000), 'High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR', *Genome Res.* Vol. 10, pp. 258–266.
3. Uhl, G.R., Liu, Q., Walther, D. *et al.* (2001), 'Polysubstance abuse-vulnerability genes: Genome scans for association, using 1,004 subjects and 1,494 single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 69, pp. 1290–1300.
4. Bansal, A., van den Boom, D., Kammerer, S. *et al.* (2002), 'Association testing by DNA pooling: An effective initial screen', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 16871–16874.
5. Gruber, J.D., Colligan, J.K. and Wolford, J.K. (2002), 'Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing', *Hum. Genet.* Vol. 110, pp. 395–401.
6. Xiao, M. and Kwok, P.Y. (2003), 'DNA analysis by fluorescence quenching detection', *Genome Res.* Vol. 13, pp. 932–939.
7. Barratt, B.J., Payne, F., Rance, H.E. *et al.* (2002), 'Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design', *Ann. Hum. Genet.* Vol. 66, pp. 393–405.
8. Jawaid, A., Bader, J.S., Purcell, S. *et al.* (2002), 'Optimal selection strategies for QTL mapping using pooled DNA samples', *Eur. J. Hum. Genet.* Vol. 10, pp. 125–132.
9. Sham, P., Bader, J.S., Craig, I. *et al.* (2002), 'DNA pooling: A tool for large-scale association studies', *Nat. Rev. Genet.* Vol. 3, pp. 862–871.
10. Barter, P.J., Brewer, Jr., H.B., Chapman, M.J. *et al.* (2003), 'Cholesteryl ester transfer protein: A novel target for raising HDL and inhibiting atherosclerosis', *Arterioscler. Thromb. Vasc. Biol.* Vol. 23, pp. 160–167.
11. Patil, N., Berno, A.J., Hinds, D.A. *et al.* (2001), 'Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21', *Science* Vol. 294, pp. 1719–1723.
12. Collins, F.S., Brooks, L.D. and Chakravarti, A. (1998), 'A DNA polymorphism discovery resource for research on human genetic variation', *Genome Res.* Vol. 8, pp. 1229–1231.
13. Zhang, K., Deng, M., Chen, T. *et al.* (2002), 'A dynamic programming algorithm for haplotype block partitioning', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 7335–7339.
14. Ballantyne, C.M., Andrews, T.C., Hsia, J.A. *et al.* (2001), 'Correlation of non-high-density lipoprotein cholesterol with apolipoprotein B: Effect of 5 hydroxymethylglutaryl coenzyme A reductase inhibitors on non-high-density lipoprotein cholesterol levels', *Am. J. Cardiol.* Vol. 88, pp. 265–269.
15. Eiriksdottir, G., Bolla, M.K., Thorsson, B. *et al.* (2001), 'The −629C > A polymorphism in the *CETP* gene does not explain the association of TaqIB polymorphism with risk and age of myocardial infarction in Icelandic men', *Atherosclerosis* Vol. 159, pp. 187–192.
16. Thompson, J.F., Lira, M.E., Durham, L.K. *et al.* (2003), 'Polymorphisms in the *CETP* gene and association with *CETP* mass and HDL levels', *Atherosclerosis* Vol. 167, pp. 195–204.
17. Knoblauch, H., Bauerfeind, A., Toliat, M.R. *et al.* (2004), 'Haplotypes and SNPs in 13 lipid-relevant genes explain most of the genetic variance in high-density lipoprotein and low-density lipoprotein cholesterol', *Hum. Mol. Genet.* Vol. 13, pp. 993–1004.
18. Hinds, D.A., Stokowski, R.P., Patil, N. *et al.* (2004), 'Matching strategies for genetic association studies in structured populations', *Am. J. Hum. Genet.* Vol. 74, pp. 317–325.
19. Chen, X., Levine, L. and Kwok, P.Y. (1999), 'Fluorescence polarization in homogeneous nucleic acid analysis', *Genome Res.* Vol. 9, pp. 492–498.
20. Knowler, W.C., Williams, R.C., Pettitt, D.J. *et al.* (1988), 'Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture', *Am. J. Hum. Genet.* Vol. 43, pp. 520–526.
21. Lander, E.S. and Schork, N.J. (1994), 'Genetic dissection of complex traits', *Science* Vol. 265, pp. 2037–2048.
22. Pritchard, J.K. and Rosenberg, N.A. (1999), 'Use of unlinked genetic markers to detect population stratification in association studies', *Am. J. Hum. Genet.* Vol. 65, pp. 220–228.
23. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), 'Inference of population structure using multilocus genotype data', *Genetics* Vol. 155, pp. 945–959.
24. Corbex, M., Poirier, O., Fumeron, F. *et al.* (2000), 'Extensive association analysis between the *CETP* gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene–environment interaction', *Genet. Epidemiol.* Vol. 19, pp. 64–80.
25. Lewontin, R.C. (1964), 'The interaction of selection and linkage. I. General considerations; heterotic models', *Genetics* Vol. 49, pp. 49–67.

26. Weir, B.S. (1996), *Genetic data analysis II*, Sinauer Associates, Sunderland, MA.
27. Ordovas, J.M., Cupples, L.A., Corella, D. *et al.* (2000), 'Association of cholesteryl ester transfer protein-TaqIB polymorphism with variations in lipoprotein subclasses and coronary heart disease risk: The Framingham study', *Arterioscler. Thromb. Vasc. Biol.* Vol. 20, pp. 1323–1329.
28. Drayna, D. and Lawn, R. (1987), 'Multiple RFLPs at the human cholesteryl ester transfer protein (CETP) locus', *Nucleic Acids Res.* Vol. 15, pp. 4698.
29. Fumeron, F., Betoulle, D., Luc, G. *et al.* (1995), 'Alcohol intake modulates the effect of a polymorphism of the cholesteryl ester transfer protein gene on plasma high density lipoprotein and the risk of myocardial infarction', *J. Clin. Invest.* Vol. 96, pp. 1664–1671.