LDL-cholesterol concentrations: a genome-wide association study

Manjinder S Sandhu, Dawn M Waterworth*, Sally L Debenham*, Eleanor Wheeler, Konstantinos Papadakis, Jing Hua Zhao, Kijoung Song, Xin Yuan, Toby Johnson, Sofie Ashford, Michael Inouye, Robert Luben, Matthew Sims, David Hadley, Wendy McArdle, Philip Barter, Y Antero Kesäniemi, Robert W Mahley, Ruth McPherson, Scott M Grundy, Wellcome Trust Case Control Consortium†, Sheila A Bingham, Kay-Tee Khaw, Ruth J F Loos, Gérard Waeber, Inês Barroso, David P Strachan, Panagiotis Deloukas, Peter Vollenweider, Nicholas J Wareham, Vincent Mooser*

Summary

Background LDL cholesterol has a causal role in the development of cardiovascular disease. Improved understanding of the biological mechanisms that underlie the metabolism and regulation of LDL cholesterol might help to identify novel therapeutic targets. We therefore did a genome-wide association study of LDL-cholesterol concentrations.

Methods We used genome-wide association data from up to 11 685 participants with measures of circulating LDL-cholesterol concentrations across five studies, including data for 293 461 autosomal single nucleotide polymorphisms (SNPs) with a minor allele frequency of 5% or more that passed our quality control criteria. We also used data from a second genome-wide array in up to 4337 participants from three of these five studies, with data for 290 140 SNPs. We did replication studies in two independent populations consisting of up to 4979 participants. Statistical approaches, including meta-analysis and linkage disequilibrium plots, were used to refine association signals; we analysed pooled data from all seven populations to determine the effect of each SNP on variations in circulating LDL-cholesterol concentrations.

Findings In our initial scan, we found two SNPs (rs599839 [$p=1{\cdot}7\times10^{-15}$] and rs4970834 [$p=3{\cdot}0\times10^{-11}$]) that showed genome-wide statistical association with LDL cholesterol at chromosomal locus 1p13.3. The second genome screen found a third statistically associated SNP at the same locus (rs646776 [$p=4{\cdot}3\times10^{-9}$]). Meta-analysis of data from all studies showed an association of SNPs rs599839 (combined $p=1{\cdot}2\times10^{-33}$) and rs646776 ($p=4{\cdot}8\times10^{-20}$) with LDL-cholesterol concentrations. SNPs rs599839 and rs646776 both explained around 1% of the variation in circulating LDL-cholesterol concentrations and were associated with about 15% of an SD change in LDL cholesterol per allele, assuming an SD of 1 mmol/L.

Interpretation We found evidence for a novel locus for LDL cholesterol on chromosome 1p13.3. These results potentially provide insight into the biological mechanisms that underlie the regulation of LDL cholesterol and might help in the discovery of novel therapeutic targets for cardiovascular disease.

Introduction

LDL cholesterol has a causal role in the development of cardiovascular disease. Indeed, experimental studies have shown the clinical efficacy of lowering concentrations of LDL cholesterol.[1] Thus, regulation of LDL cholesterol represents a fundamental target for devising additional interventional strategies to reduce the risk of cardiovascular disease. In this context, understanding the biological mechanisms that underlie metabolism and regulation of LDL cholesterol might help to identify novel therapeutic targets.

Variation in LDL-cholesterol concentrations is a polygenic trait.[2–4] Integration of genome-wide technologies and epidemiological approaches could help to identify novel genetic determinants of variation in LDL-cholesterol concentrations, providing new insights into the metabolism and regulation of LDL cholesterol.[5,6] We therefore did a genome-wide association study on 11 685 participants with measures of circulating LDL-cholesterol concentrations across five studies. To validate these associations, we also did replication studies in independent populations.

Methods

Participants

Data were gathered from five groups of individuals: two subcohorts of the EPIC-Norfolk study, the 1958 British birth cohort, the CoLaus study, and the Genetic Epidemiology of Metabolic Syndrome study.

The EPIC-Norfolk study is a population-based cohort study of 25 663 white European men and women aged 39–79 years recruited in Norfolk, UK, between 1993 and 1997.[7] We examined a subcohort that consisted of 2566 individuals who were randomly selected from the total cohort using a random selection algorithm. Serum total cholesterol, HDL cholesterol, and triglycerides were measured in fresh samples with the RA-1000 analyser (Bayer Diagnostics, Basingstoke, UK). LDL-cholesterol concentrations were calculated with the Friedewald formula.[8] The Norwich local research ethics committee granted ethical approval for the study. All participants gave written informed consent.

The EPIC-Norfolk obese set is a case series also derived from the EPIC-Norfolk cohort, consisting of

Lancet 2008; 371: 483–91

See **Comment** page 450

*Contributed equally

†Members listed at end of paper

Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Cambridge, UK (M S Sandhu PhD, S L Debenham PhD, S Ashford BSc, R Luben BSc, Prof K-T Khaw FRCP); **MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK** (M S Sandhu, S L Debenham, J H Zhao PhD, S Ashford, M Sims BSc, R J F Loos PhD, Prof N J Wareham FRCP); **Division of Medical Genetics/Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, King of Prussia, PA, USA** (D M Waterworth PhD, K Song PhD, X Yuan PhD, V Mooser MD); **Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK** (E Wheeler PhD, M Inouye MSc, I Barroso PhD, P Deloukas PhD); **Department of Medical Genetics and University Institute of Social and Preventive Medicine, University of Lausanne and Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland** (T Johnson PhD); **Swiss Institute of Bioinformatics, Switzerland** (T Johnson); **Division of Community Health Sciences, St George's, University of London, London, UK** (K Papadakis MSc, D Hadley MSc, Prof D P Strachan MD); **Avon Longitudinal Study of Parents and Children, University of Bristol, Bristol, UK** (W McArdle PhD); **Heart Research Institute, Camperdown, Sydney, NSW, Australia** (Prof P Barter PhD);

Department of Internal Medicine and Biocenter Oulu,University of Oulu, Oulu, Finland (Prof Y A Kesäniemi PhD); **Gladstone Institute of Neurological Disease and Gladstone Institute of Cardiovascular Disease, San Francisco, CA, USA** (Prof R W Mahley PhD); **Division of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada** (Prof R McPherson FRCP); **Center for Human Nutrition, Department of Clinical Nutrition, University of Texas Southwestern Medical Center, Dallas, TX, USA** (Prof S M Grundy PhD); **MRC Dunn Human Nutrition Unit, Cambridge, UK** (Prof S A Bingham PhD); **and Department of Internal Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland** (Prof G Waeber MD, P Vollenweider MD)

Correspondence to:
Dr Manjinder S Sandhu, Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Cambridge CB1 8RN, UK
manj.sandhu@srl.cam.ac.uk

1685 individuals with obesity (body-mass index [BMI] ≥30 kg/m²). These cases were selected independently from the EPIC-Norfolk subcohort. Of these, 1284 cases were non-overlapping and used as a further study set. Serum total cholesterol, HDL cholesterol, and triglycerides were measured in fresh samples with the RA-1000 analyser, and LDL-cholesterol concentrations were calculated with the Friedewald formula.[8] The Norwich local research ethics committee granted ethical approval for the study. All participants gave written informed consent.

The third study was the 1958 British birth cohort, a national population sample followed up periodically from birth to age 44–45 years, when a DNA bank was established as a national reference series for case-control studies.[9] A geographically representative subsample of 1480 participants who were selected as controls for the Wellcome Trust Case-Control Consortium genome-wide-association studies[10] were included in this analysis. Triglycerides, serum total cholesterol, and HDL cholesterol were measured in non-fasting serum with the Olympus model AU640 autoanalyser (Olympus Inc, Center Valley, PA, USA) by a clinical biochemistry laboratory. The concentration of LDL cholesterol was derived by the Friedewald formula.[8] All participants included in this analysis gave written informed consent for the use of their DNA for non-commercial medical research purposes. Field protocols, informed consent, and this within-cohort genetic association analysis were approved by the South East NHS Multi-Centre Research Ethics Committee.

Participants in the CoLaus (Cohorte Lausannoise) study were randomly selected from 56 694 individuals aged 35–75 years who were permanent residents of Lausanne, Switzerland.[11] Recruitment took place between April, 2003, and March, 2006, with 6186 individuals participating in the study. Of those invited to take part, 41% actually participated. Only white European individuals (ie, individuals for whom the four grandparents were white European) were included in the study. Participants provided a detailed health questionnaire and underwent a physical examination, including measurements of anthropometric variables. Participants donated blood after a 12-h fasting period for clinical chemistry and genetic analyses. Nuclear DNA was extracted from whole blood for whole genome scan analysis. Clinical chemistry assays were done by a clinical laboratory at the Centre Hospitalier Universitaire Vaudois on fresh blood samples on a Modular P apparatus (Roche Diagnostics, Basel, Switzerland) within 2 h of blood collection. Total cholesterol was assessed by CHOD-PAP (Roche Diagnostics, Basel, Switzerland; maximum inter-batch CV 1·6%; maximum intra-batch CV 1·7%) and HDL cholesterol by CHOD-PAP, polyethylene glycol, and cyclodextrin (maximum inter-batch CV 3·6%; maximum intra-batch CV 0·9%). LDL cholesterol was calculated with the Friedewald formula.[8] The study was approved by the ethical committee of the faculty of medicine of Lausanne. The study was sponsored in part by GlaxoSmithKline, and all participants were duly informed about this sponsorship, and consented for the use of biological samples and data by GlaxoSmithKline and its subsidiaries.

The study population of the Genetic Epidemiology of Metabolic Syndrome (GEMS) study consisted of dyslipidaemic cases (age 20–65 years, n=1025) matched with normolipidaemic controls (n=1008) by sex and recruitment site.[12] Detailed information on the GEMS study design, sampling frame, and recruitment procedures has been published.[12] Serum triglycerides were measured enzymatically after hydrolysation to glycerol (Hitachi 704 analyser; Hitachi, Tokyo, Japan). HDL cholesterol was measured after the precipitation of other lipoproteins with a heparin-manganese chloride mixture (Hitachi 704 analyser). Dyslipidaemic participants were defined as those with triglycerides above the 75th percentile and HDL cholesterol below the 25th percentile, on the basis of age, sex, and country-specific distributions. Normolipidaemic controls (triglycerides <50th percentile, HDL cholesterol >50th percentile, and BMI >25 kg/m²), aged over 40 years were ascertained at the same time. Blood samples were collected after a 12-h fast and LDL-cholesterol concentration was calculated with the Friedewald formula.[8] The study was sponsored in part by GlaxoSmithKline, and all participants were duly informed about this sponsorship, and consented for the use of biological samples and data by GlaxoSmithKline and its subsidiaries; the study was approved by the local ethics committees.

For confirmation of our results, we also included another subset from EPIC-Norfolk, consisting of up to 3339 participants who did not overlap with the EPIC-Norfolk subcohort or obese set. These individuals were sequentially selected from those with DNA available for genotyping. We also used 1697 participants with DNA available for genotyping from the Ely study[13] as another replication cohort—a population of white European men and women aged 35–79 years without diagnosed diabetes. This study is a population-based cohort study of the cause and pathogenesis of type 2 diabetes and related metabolic disorders in the UK. The Cambridge research ethics committee approved the study. All participants gave informed consent.

Selected study characteristics of all study populations are provided in webtable 1.

See Online for webtable 1

Procedures

Participants were genotyped with the Affymetrix GeneChip Human Mapping 500K array set (Santa Clara, CA, USA). To optimise data quality and statistical analyses, sample and single nucleotide polymorphism (SNP) quality control criteria and statistical analysis of LDL cholesterol were done within each study, independent of the other studies. For the initial genome-wide association screen, analyses were also done within study. Thus, the relevant linear regression analyses were

optimised on the basis of the specific characteristics of the individual studies.

For the EPIC-Norfolk subcohort and the EPIC-Norfolk obese set, SNP genotyping was done at the Affymetrix services laboratory (San Francisco, CA, USA). Genotypes were obtained by the BRLMM algorithm clustered by plate.[14] 2566 participants were genotyped in the EPIC-Norfolk subcohort and 1284 in the obese set. Individuals were excluded by use of the following sample quality control criteria: proportion of all genotypes called was less than 94%; heterozygosity was less than 23% or more than 30%; if there was more than 5·0% discordance in SNP pairs with r^2=1 in HapMap; ethnic outliers; related individuals (>70% concordance with another sample); and duplicates (concordance with another DNA was >99%). These exclusions left 2417 participants in the subcohort and 1135 in the obese set. We then applied the following SNP quality control criteria. SNPs were dropped if the proportion of genotypes called was 90% or less, if they were not in Hardy-Weinberg equilibrium ($p<1\times10^{-6}$), and if they had a minor allele frequency (MAF) of 5% or less. We also restricted this analysis to autosomal SNPs. Therefore, there were 344837 SNPs analysed in this genome-wide association scan for the EPIC-Norfolk subcohort. We also used the same subset of SNPs for analysis of the obese set. After sample and SNP quality control, 2269 individuals in the first subcohort and 1009 of those in the obese set had a measure of LDL cholesterol. LDL-cholesterol concentrations showed a near normal distribution in both studies. We therefore used untransformed LDL-cholesterol data. Linear regression analysis was used to assess the association between each SNP and LDL-cholesterol concentration with an additive (per allele) model (1 df) by use of PLINK version 1.0. No covariables were included in these analyses.

For the 1958 British birth cohort, genotypes were measured with the Affymetrix GeneChip 500K array and processed by the CHIAMO algorithm.[10] After sample quality control checks for contamination, non-white European identity, overall heterozygosity, relatedness, low proportion of called genotypes, and evidence of non-white European ancestry, 1480 individuals were available with genome-wide association data. SNPs were excluded from the analysis because of missing data, departures from Hardy-Weinberg equilibrium and other metrics.[10] These exclusions left 461986 SNPs for analysis. 1375 participants had data for LDL-cholesterol concentrations. Linear regression models were used to test the additive (per allele) effect of the minor allele at each locus on untransformed LDL-cholesterol concentrations. Covariables included in this analysis included sex and study-specific factors. Statistical analysis of the measured SNPs was done with Stata version 8.1.

For the CoLaus study, genotyping was done with Affymetrix GeneChip Human Mapping 500K array set according to the Affymetrix protocol. Genotypes were obtained with the BRLMM algorithm. Participants were removed from the analysis on the basis of the following sample quality control criteria: any participant whose sex was inconsistent with genetic data from X-linked SNPs; the proportion of genotypes called was less than 90%; having inconsistent genotypes when compared with duplicate samples. 5636 participants remained after sample quality control exclusions. We then applied SNP exclusions with the following criteria: SNPs that were monomorphic among all samples; SNPs with genotypes on less than 95% participants; SNPs that were out of Hardy-Weinberg equilibrium ($p<1\cdot0\times10^{-7}$). After these quality control procedures, 370697 SNPs remained for analysis. 5367 participants had a measure of LDL cholesterol. LDL-cholesterol concentrations showed some evidence of a non-normal distribution. Therefore, all regression analyses were done on natural log-transformed LDL-cholesterol concentrations, which showed a near normal distribution. We used linear regression analysis with an additive model adjusted for age, sex, and geographic origin. Analyses were done with PLINK version 1.0.

For the GEMS study, genotyping was done with the Affymetrix GeneChip Human Mapping 500K array and the BRLMM calling algorithm. We excluded individuals on the basis of the same sample quality control criteria as for the CoLaus study. After sample quality control, 1847 participants from the original sample remained. We then did a SNP quality assessment, excluding SNPs that were monomorphic, those that were out of Hardy-Weinberg equilibrium ($p<1\cdot0\times10^{-7}$) or if the proportion of genotypes called for each SNP was less than 95%. After SNP quality control, 359052 SNPs were available for analysis. 1665 participants had a measure of LDL cholesterol. Again, LDL-cholesterol concentrations seemed to have a non-normal distribution. We therefore did linear regression analysis with natural log-transformed LDL-cholesterol data and an additive model with adjustment for age, sex, study site, and dyslipidaemia status. Analyses were done with PLINK version 1.0.

To provide additional support for association signals, we examined data from the three of the five studies with a separate genotyping platform and SNP array. Participants from the EPIC-Norfolk subcohort and the EPIC-Norfolk obese set were also genotyped with the Infinium HumanHap300 SNP chip (Illumina, San Diego, CA, USA), containing 317503 tagging SNPs derived from phase I of the International HapMap project. Of the SNPs assayed on these chips, we excluded SNPs if the proportion genotyped was 90% or less, if the MAF was 5% or less, and if the genotype distribution was out of Hardy-Weinberg equilibrium ($p<1\cdot0\times10^{-6}$). Participants from the 1958 British birth cohort were genotyped with the Infinium HumanHap550 SNP chip (Illumina). Details of the sample and SNP quality control criteria have been published elsewhere.[15] We did within-study analyses with an approach identical to the

analyses of Affymetrix data. All Illumina genotyping was done at the Wellcome Trust Sanger Institute (Hinxton, Cambridgeshire, UK).

The EPIC-Norfolk replication set and the Ely study were genotyped with custom TaqMan SNP assays (Applied Biosystems, Warrington, UK) at Strangeways research laboratory (University of Cambridge, Cambridge, UK).

See Online for webtable 2

Statistical analysis

To increase statistical precision in the initial genome-wide association analysis with Affymetrix data, we meta-analysed summary data from each of the five studies by use of a fixed effects model and inverse-variance weighted averages of β coefficients with Stata version 8.2. We therefore obtained a combined estimate of the overall β coefficient and its SE. Between-study heterogeneity was assessed with the χ^2 test. To optimise data quality and statistical efficiency, we only analysed SNPs that passed sample and SNP quality control criteria in each of the five studies and that had a measure of association (β coefficient and SE) in all five studies.

We attempted to reduce the effect of population stratification (confounding) by use of appropriate epidemiological design and statistical analysis, with ethnically homogeneous populations within each study. This approach also included adjusting for possible population stratification within study by use of geographical covariables, where appropriate. In the same context, all analyses were done conditioning on study. For each study, we also used quantile-quantile plots of the observed and expected distributions of p values to assess whether there was any evidence of distortion of the observed distribution from the null (data not shown). We also calculated an inflation factor (λ) for each study,[16] which was estimated from the mean of the χ^2 tests generated on all SNPs that were tested. On the basis of data for the 293 461 tested SNPs included in the meta-analysis, the distribution of the test statistics closely followed the null distribution for each study. Accordingly, the inflation factor was close to 1 for each study (webtable 2), suggesting that the observed associations are unlikely to be the result of population stratification or other artefacts. We also used this approach to calculate an inflation factor for the combined data. Dividing the χ^2 statistics for SNPs reaching genome-wide statistical association by this inflation factor did not alter the interpretation of these findings (data not shown).

Meta-analysis of data from Illumina assays was done in much the same way as for Affymetrix data. Quantile-quantile plots of association within study indicated that the data followed the null distribution. By use of the method of genomic control, we again found that the inflation factor was close to 1 for each study (webtable 2), suggesting that the observed associations are unlikely to be the result of population stratification.

We used pairwise correlation (r^2) to assess the extent of linkage disequilibrium between co-located SNPs. On the basis of linear regression analyses, we then used likelihood ratio tests to assess whether statistically associated SNPs independently contributed to the variation in LDL-cholesterol concentrations and to determine the source of any association signal. For these analyses, individual participant data were available for only four studies (no data were available for the 1958 British birth cohort). Specifically, we compared the log likelihood of a nested model (2 df) with that of the full model (3 df) by consecutively adding extra SNPs (in a log

	Chromosome	Position*	Nearest locus	Minor allele†	Frequency†	Pooled β coefficient (SE)‡	Combined p value	p value for heterogeneity	Rank
rs4420638	19	50114786	APOC1	G	0·18	0·06 (0·01)	$1·2\times10^{-20}$	$2·8\times10^{-9}$	1
rs599839	1	109623689	PSRC1	G	0·21	−0·05 (0·01)	$1·7\times10^{-15}$	$2·0\times10^{-5}$	2
rs4970834	1	109616403	CELSR2	T	0·19	−0·04 (0·01)	$3·0\times10^{-11}$	0·01	3
rs562338	2	21141826	APOB	T	0·20	−0·04 (0·01)	$1·4\times10^{-9}$	$3·1\times10^{-5}$	4
rs7575840	2	21126995	APOB	T	0·34	0·03 (0·01)	$1·9\times10^{-9}$	$4·8\times10^{-4}$	5
rs478442	2	21252721	APOB	G	0·21	−0·03 (0·01)	$8·1\times10^{-9}$	$4·4\times10^{-4}$	6
rs4591370	2	21237247	APOB	A	0·21	−0·03 (0·01)	$8·2\times10^{-9}$	$2·5\times10^{-4}$	7
rs4560142	2	21237222	APOB	C	0·21	−0·03 (0·01)	$8·3\times10^{-9}$	$4·4\times10^{-4}$	8
rs576203	2	21247128	APOB	A	0·21	−0·03 (0·01)	$9·0\times10^{-9}$	$3·0\times10^{-4}$	9
rs506585	2	21250687	APOB	G	0·21	−0·03 (0·01)	$1·0\times10^{-8}$	$3·9\times10^{-4}$	10
rs488507	2	21247194	APOB	G	0·22	−0·03 (0·01)	$2·0\times10^{-8}$	$1·3\times10^{-3}$	11
rs538928	2	21242524	APOB	A	0·20	−0·03 (0·01)	$2·7\times10^{-8}$	$6·8\times10^{-4}$	12
rs10402271	19	50021054	BCAM	G	0·33	0·03 (0·01)	$4·1\times10^{-8}$	0·02	13
rs693	2	21085700	APOB	C	0·47	−0·03 (0·01)	$4·4\times10^{-8}$	0·02	14

*On basis of NCBI Build 36·2. †On basis of EPIC-Norfolk sub-cohort: minor allele corresponds to forward strand of NCBI Build 36·2. ‡β coefficients represent the change in LDL-cholesterol concentration per additional minor allele.

Table 1: **Statistical associations ($p<1·0\times10^{-7}$) between Affymetrix SNPs and circulating concentrations of LDL cholesterol in a genome-wide meta-analysis of five study populations consisting of up to 11 685 participants**

	Chromosome	Position*	Nearest locus	Minor allele†	Frequency†	Pooled β coefficient (SE)‡	Combined p value	p value for heterogeneity	Rank
rs2075650	19	50087459	TOMM40	G	0·13	0·23 (0·03)	$7·1\times10^{-14}$	0·15	1
rs4803750	19	49939467	BCL3	G	0·07	−0·28 (0·04)	$2·4\times10^{-11}$	0·14	2
rs646776	1	109620053	CELSR2	G	0·21	−0·16 (0·03)	$4·3\times10^{-9}$	0·70	3
rs1713222	2	21124828	APOB	T	0·16	−0·17 (0·03)	$1·0\times10^{-8}$	0·56	4
rs2228671	19	11071912	LDLR	T	0·12	−0·18 (0·03)	$1·1\times10^{-8}$	0·50	5
rs11668477	19	11056030	LDLR	G	0·20	−0·15 (0·03)	$1·5\times10^{-8}$	0·95	6
rs4605275	19	50030333	BCAM	T	0·31	−0·13 (0·02)	$4·7\times10^{-8}$	0·74	7

*On basis of NCBI Build 36·2. †On basis of EPIC-Norfolk subcohort; minor allele corresponds to forward strand of NCBI Build 36·2. ‡β coefficients represent the change in LDL-cholesterol concentration per additional minor allele.

Table 2: **Statistical associations ($p<1·0\times10^{-7}$) between Illumina SNPs and circulating concentrations of LDL cholesterol in a genome-wide meta-analysis of three UK study populations consisting of up to 4337 participants**

additive form) to a model containing the SNP with the strongest statistical signal from the genome-wide screen (general inheritance [2 df] form). We also did a reciprocal analysis, adding the SNP with the strongest statistical signal from the genome-wide screen (1 df form) to a model containing other SNPs showing statistical association (2 df form). All analyses were done with Stata version 8.2.

We then generated a linkage disequilibrium plot with Haploview.[17] Linkage disequilibrium blocks are delineated by black lines and defined with the method of Gabriel and colleagues.[18] To provide a more detailed assessment of chromosomal regions showing statistical association, and to further refine the location of any association signal, we imputed SNPs on the basis of HapMap phase II data with IMPUTE.[19] We used information on SNP genotypes in our studies and HapMap II data to statistically predict all SNP genotypes in a chromosomal region for all individuals. For these imputed data, association analysis was done with SNPTEST (with the full posterior probability genotype distribution) for the imputed genotypes and LDL-cholesterol data for each study separately (adding in relevant covariables). β coefficients were combined as before. Only SNPs with a MAF of 1% or more and with a posterior-probability score more than 0·90 were considered for these imputed association analyses.[10] Imputed data were not available for the GEMS study.

To increase comparability between studies, we reanalysed data from CoLaus and GEMS where we had individual participant data with untransformed LDL-cholesterol data and with no covariable adjustment. We then did a meta-analysis of all studies, including the replication studies and untransformed data from CoLaus and GEMS. As before, we obtained β coefficients and SE from each study and used a fixed effects model and inverse-variance weighting to obtain a combined estimate of the overall β coefficient and its SE. Heterogeneity between studies was assessed with χ^2 tests. Finally, we did a pooled analysis (conditioning on study) of all studies in which we had individual participant data. This analysis provided an estimate of the magnitude of the relation between these SNPs and LDL-cholesterol concentrations in a comparable way, and also provided a measure of the variation in circulating LDL-cholesterol concentrations explained by each SNP. We also used multivariable linear regression analysis to examine whether age and sex affected these associations. By use of these pooled data, we also contextualised our results as a proportion of a SD change in LDL-cholesterol concentrations. All analyses were done with Stata version 8.2. For a downloadable file of all results, including additional analyses, see http://www.srl.cam.ac.uk/gecd/.

Role of the funding source

This study was funded by UK Medical Research Council, Wellcome Trust, British Heart Foundation, and GlaxoSmithKline. Employees of GlaxoSmithKline contributed to the study design, data analysis, data interpretation, and writing of the report. Other sponsors of the study had no role in data collection, the study design, data analysis, data interpretation, or writing of the report. The corresponding author had full access to the data and final responsibility for the decision to submit for publication.

Results

Data for 293 461 autosomal SNPs identified with Affymetrix were available for analysis in up to 11 685 individuals from five studies. 14 SNPs were statistically associated with circulating LDL-cholesterol concentrations at the genome-wide level ($p<1·0\times10^{-7}$; table 1). The 14 SNPs showed directionally consistent signals in all five studies (webtable 3). These SNPs were broadly confined to three distinct genomic regions (table 1), two of which were close to known loci involved in LDL-cholesterol metabolism (including those encompassing the *APOB* and *APOE* genes).[2,20,21] Notably, two SNPs—rs599839 ($p=1·7\times10^{-15}$) and rs4970834 ($p=3·0\times10^{-11}$)—showed evidence for statistical association and were both located at chromosomal region 1p13.3. For

See Online for webtable 3

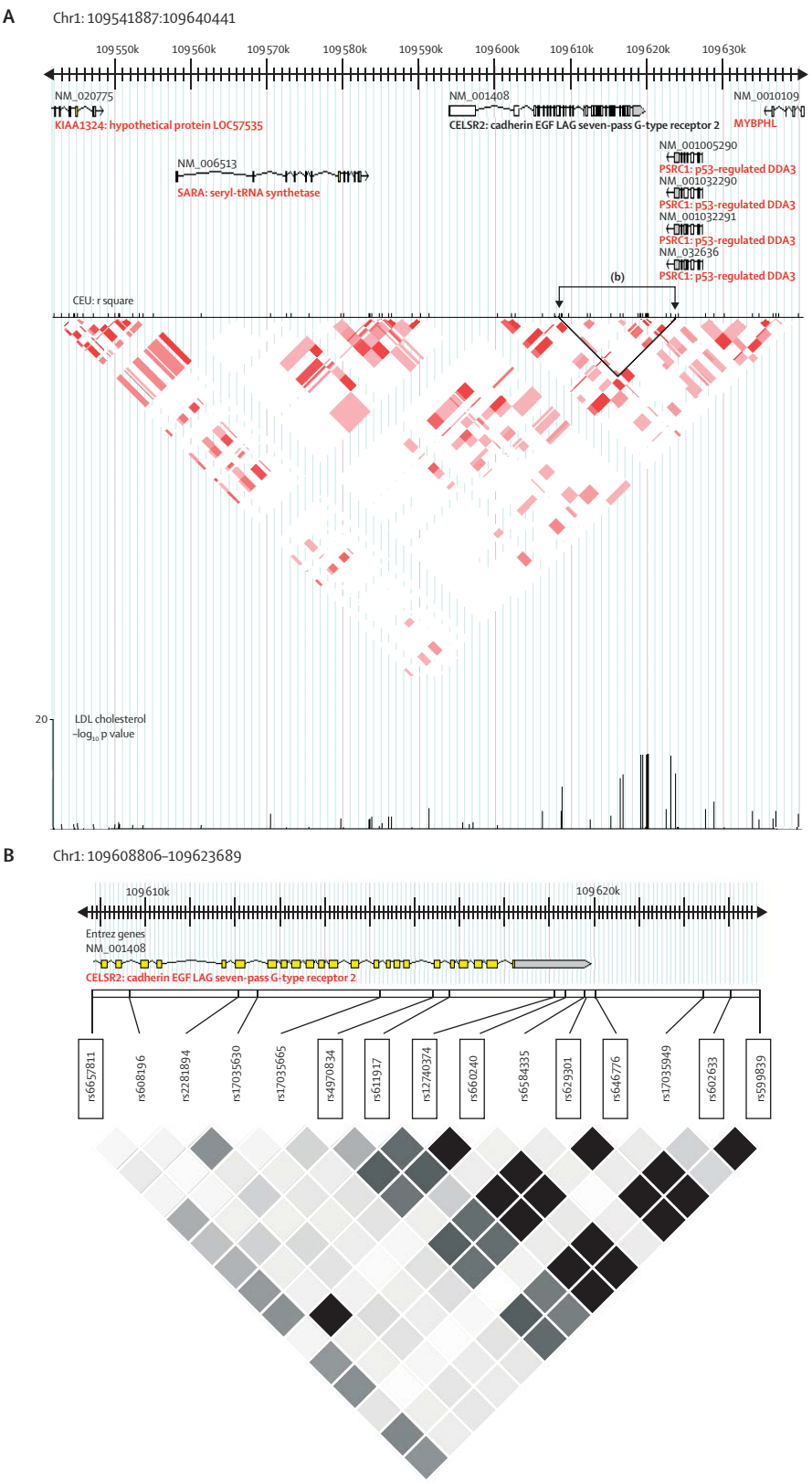

Figure 1: **Linkage disequilibrium plot**

(A) Plot of 98 kb genomic region aligned with association signals for imputed SNPs. Positions of genes, SNPs genotyped in HapMap, and linkage disequilibrium among SNPs (r^2 is shown). r^2 values of 1·0 are depicted by red diamonds, intermediate r^2 values are represented in pink, and r^2 values of 0 as white. Aligned underneath the linkage disequilibrium plot is a graph showing the association signal for each of the 71 SNPs which could be imputed from our data. The plot was generated with HapMap (release 22/phase II Apr 07, NCBI B36 assembly, dbSNP build 126, [CEPH Utah trios], chr1 co-ordinates 109541887-109640441). This plot illustrates that the strongest association signals are localised to a 14 kb region shown in detail in (B). Imputed SNPs that were statistically associated with circulating concentrations of LDL cholesterol at the genome-wide level ($p<1{\cdot}0\times10^{-7}$) are boxed.

	Genomic position*	Position relative to *CELSR2* gene	Minor allele†	Frequency†	Pooled β-coefficient (SE)	Combined p value	p value for heterogeneity
rs646776	109620053	3′ (intergenic)	G	0·21	–0·13 (0·02)	$3·0\times10^{-14}$	0·16
rs629301	109619829	3′ UTR	C	0·21	–0·13 (0·02)	$3·1\times10^{-14}$	0·15
rs12740374	109619113	3′ UTR	T	0·21	–0·13 (0·02)	$3·2\times10^{-14}$	0·15
rs660240	109619361	3′ UTR	A	0·21	–0·14 (0·02)	$3·8\times10^{-14}$	0·17
rs602633	109623034	3′ intergenic	A	0·22	–0·13 (0·02)	$5·7\times10^{-14}$	0·18
rs599839	109623689	3′ (intergenic)	G	0·19	–0·12 (0·02)	$7·8\times10^{-11}$	0·45
rs611917	109616775	Intronic	C	0·28	–0·11 (0·02)	$1·5\times10^{-10}$	0·05
rs4970834	109616403	Intronic	T	0·17	–0·12 (0·02)	$6·7\times10^{-10}$	0·41
rs6657811	109608806	Intronic	T	0·12	–0·13 (0·02)	$2·0\times10^{-8}$	0·04

UTR=untranslated region. *On basis of NCBI Build 36·2. †On basis of EPIC-Norfolk subcohort; minor allele corresponds to forward strand of NCBI Build 36·2. ‡β coefficients represent the change in LDL-cholesterol concentration per additional minor allele.

Table 3: **Imputed SNPs showing genome-wide statistical association ($p<1·0\times10^{-7}$) with circulating concentrations of LDL cholesterol; meta-analysis of four study populations consisting of up to 9988 participants**

these SNPs, there was no material evidence for heterogeneity between studies after adjustment for multiple testing (table 1).

Data for 290 140 SNPs identified with Illumina were available for analysis across the three studies that were assessed with these chips. For these analyses, we had up to 4337 participants with a measure of LDL-cholesterol concentration. Seven SNPs were statistically associated with circulating LDL-cholesterol concentrations at the genome-wide level ($p<1·0\times10^{-7}$; table 2 and webtable 4). Six of these SNPs were located in genomic regions previously linked to LDL-cholesterol metabolism. However, we also found another SNP located at chromosomal region 1p13.3 (rs646776; $p=4·3\times10^{-9}$).

Linkage disequilibrium plots of the three SNPs located at 1p13.3 implicated a region spanning several genes (webfigure). The strongest statistically associated SNP (rs599839) lay 3′ to the *CELSR2* and *PSRC1* genes (the two genes are in a tail-to-tail orientation) in a 98 kb region of fragmented linkage disequilibrium. This region contained several recombination hotspots and was situated between two blocks of strong linkage disequilibrium (webfigure). SNP rs4970834 also lay in this region and in our studies was correlated with SNP rs599839 ($r^2=0·79$; webtable 5). SNP rs646776 was also colocalised with these SNPs and was highly correlated ($r^2=0·94$) with SNP rs599839 (webtable 5). For all three SNPs, the minor allele, with a frequency of around 19–21%, was associated with lower LDL-cholesterol concentrations (table 1 and table 2).

Likelihood ratio tests of up to 10 310 participants showed that, for the Affymetrix SNPs, assuming that SNP rs599839 was the causal variant or in near complete linkage disequilibrium with the causal variant(s), inclusion of SNP rs599839 as a covariable explained the other observed SNP associations in this region (webtable 6). In an exploratory and equivalent analysis on a small subset of samples with both Affymetrix and Illumina SNP data (up to 3007 participants), we found that both SNPs rs599839 and rs646776, which are highly correlated, equally explained the association signals in this region (webtable 6). Thus, when conditioning on SNP rs599389, our results indicated that the three statistically associated SNPs might be characterising identical genetic variant(s) in this region.

Imputation of all SNPs with a MAF of 1% or more from our Affymetrix array and HapMap II data for this 98 kb region for 9988 participants allowed us to assess whether additional association signals might be present in this region. On the basis of these imputed data, the strongest evidence for association was found for SNP rs646776 ($p=3·0\times10^{-14}$; figure 1 and table 3). Indeed, within this 98 kb region, the strongest association signals from both imputed and genotyped data were localised to a 14 kb region containing a group of seven highly correlated SNPs that included the 3′ untranslated region (UTR) of the *CELSR2* gene (figure 1 and table 3). By use of data from HapMap, we found that SNPs rs599839 and rs646776 tag six of these seven SNPs with an r^2 between 0·96 and 1·00. However, in view of the strong linkage disequilibrium, the specific source of the association signal is unlikely to be reliably differentiated between these SNPs in our data.

See Online for webtables 4, 5, 6, and 7

See Online for webfigure

To validate associations found in our genome-wide association screens and imputational analysis, we genotyped rs599839 and rs646776 in the two replication cohorts (webtable 7). There was again evidence of statistical association in each of these studies (webtable 7), thus corroborating our imputed results.

Lastly, we conducted a meta-analysis and pooled analysis of all available studies by use of a comparable analytical approach. Meta-analysis of data for up to 16 571 participants showed evidence for statistical association for SNP rs599839 with LDL-cholesterol concentrations ($p=1·2\times10^{-33}$; webtable 7). On the basis of data for up to 9282 participants, we found evidence for statistical association for SNP rs646776 with LDL-cholesterol concentrations ($p=4·8\times10^{-20}$). These

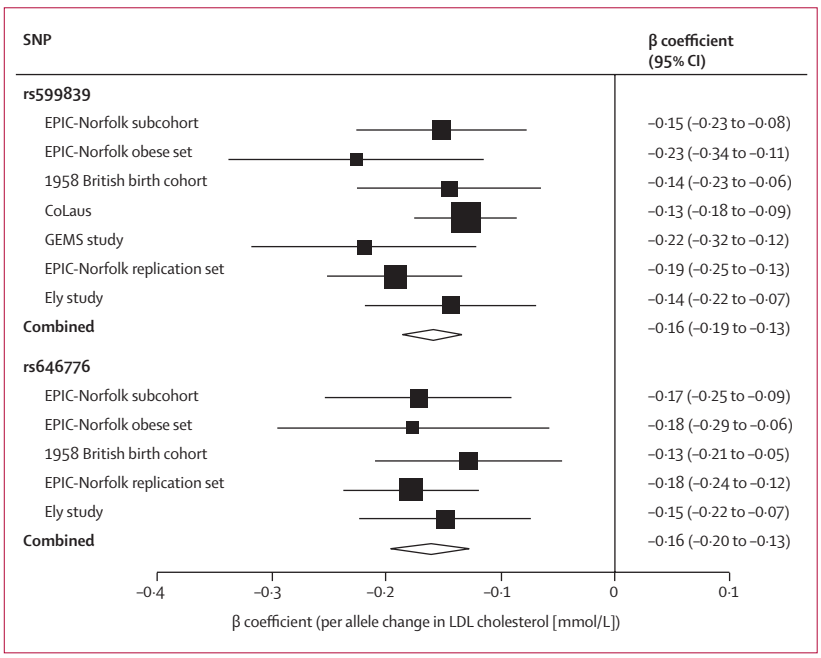

Figure 2: **Association between SNPs at the 1p13.3 locus and circulating concentrations of LDL cholesterol**
For individual studies β coefficients are depicted by black boxes and spanned by 95% CI. Diamonds represent overall β coefficients for each SNP and the width of the diamonds delineate their 95% CI. Corresponding values for each independent study and for the overall estimate are given to the right of the plot.

associations were directionally consistent across all studies (figure 2), with no heterogeneity between studies (p=0·43 for SNP rs599839 and p=0·88 for rs646776).

A pooled analysis of all studies in which we had individual participant data, which consisted of 15 196 individuals for SNP rs599839 and 7952 individuals for SNP rs646776, suggested that SNPs rs599839 and rs646776 both explained around 1% of the variation in circulating LDL-cholesterol concentrations and were associated with about 15% of an SD change in LDL-cholesterol concentrations per allele (figure 2). Further adjustment for age and sex did not alter these findings (data not shown).

Discussion

Our data provide evidence for a locus for LDL cholesterol at chromosome region 1p13.3. This locus has not previously been related to lipid metabolism.[4,20] These results could provide insight into the biological mechanisms that underlie the regulation of LDL-cholesterol concentrations and might help to identify new therapeutic targets for cardiovascular disease. The magnitude of the association was consistent across the studies we examined, and showed independent evidence for statistical association in each study.

Examination of the publicly available Affymetrix 500K results from the Diabetes Genetics Initiative genome-wide association scan of LDL-cholesterol concentrations[22] suggests that rs599839 and rs4970834—the two most strongly associated SNPs from our Affymetrix array—also showed clear evidence for statistical association ($p=9\cdot0\times10^{-8}$ and $p=1\cdot2\times10^{-4}$, respectively) with LDL-cholesterol concentrations, providing independent confirmation of our findings.

Our results are unlikely to be artifacts. We used several genotyping technologies, independent replication, and stringent statistical criteria to define our associations.[23] The consistency of the association across heterogeneous populations also argues against a false positive association. By contrast, random error in the measurement of LDL-cholesterol concentrations might have lead to an underestimation of the magnitude of the association between these genetic variants and LDL-cholesterol concentrations. However, it is likely that our study does not have the statistical power to detect other novel genetic determinants with smaller effects on LDL-cholesterol concentrations. Even larger scale studies will be required to detect these associations. Consistent with a causal link between LDL cholesterol and risk of coronary artery disease, the locus that we identified has also shown statistical association with risk of coronary heart disease in other genome-wide association studies.[24] This association is directionally consistent with our data. Specifically, our data show that minor allele carriers of these genetic variants, who made up around 20% of our populations, have lower circulating LDL-cholesterol concentrations. In keeping with this finding, a genome-wide association of coronary heart disease found that individuals with these alleles have a lower risk of developing coronary artery disease than do individuals who are homozygous for the major allele.[24] In this context, our study also shows that, in addition to discovering new genetic determinants of quantitative traits, genome-wide association studies of quantitative risk factors can provide a research framework to determine the mechanism underlying association signals for relevant disease susceptibility genes.

Genetic variants at this locus explained around 1% of the variation in LDL-cholesterol concentrations, and were associated with about 15% of an SD change in LDL-cholesterol concentrations per allele, on the basis of a SD of around 1 mmol/L. With caveats, these variants might therefore have use as genetic tools for causal inference in Mendelian randomisation studies of cardiovascular disease.[25]

Up to now, genetic variation at the apolipoprotein E and B genes, the LDL receptor gene, and variation at the gene encoding proprotein convertase subtilisin/kexin type 9 (*PCSK9*) has been consistently shown to affect LDL-cholesterol concentrations.[2,20,21,26–28] Mutations in these genes are also causes of familial hypercholesterolaemia.[2,27,29] Our data indicate that association signals between genetic variants at the 1p13.3 locus and LDL-cholesterol concentrations might be localised to the 3′ UTR of the *CELSR2* gene—the cadherin EGF LAG seven-pass G-type receptor 2. Further genetic epidemiological studies could help clarify the source of

the association signal or the causal variant(s). However, because of the strong correlation between statistically associated SNPs at this gene in European populations, studies of populations with greater genetic diversity might be required to help resolve these association signals. The biological role of the *CELSR2* gene is unknown. Functional studies and examination of genetic mutations in this region might help clarify the role of proteins encoded by this genomic region in lipid metabolism and disorders, including familial hypercholesterolaemia.

Contributors

MSS, SLD, EW, DMW, KP, DH, DS, JHZ, KS, XY, TJ, and RL did the statistical analyses. PD and MI coordinated the genome-wide association genotyping and bioinformatics for the EPIC-Norfolk and 1958 British birth cohort studies. DMW coordinated the genome-wide association genotyping and bioinformatics for the CoLaus and GEMS studies. MS and SA contributed to replication genotyping. MSS wrote the report. All authors contributed to the study design, analysis, and interpretation of the data.

Wellcome Trust Case Control Consortium

Management committee: Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, John A Todd, Peter Donnelly (Chair).

Conflict of interest statement

DMW, KS, XY, VM are employees of GlaxoSmithKline. MSS and RM have received research funding from GlaxoSmithKline. SMG has consulted for GlaxoSmithKline. AK has received research funding from, has provided CME on behalf of, and has acted as a consultant to: AstraZeneca, Laboratories Fournier, Merck/Schering Plough, Novartis, Pfizer and Sanofi-Aventis, and also owns some Orion-Pharma stocks.

Acknowledgments

We acknowledge the support of the UK Medical Research Council, Wellcome Trust, British Heart Foundation (BHF), European Commission, and GlaxoSmithKline. Specifically, we acknowledge use of genotype data from the 1958 British birth cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. IB and EW acknowledge support from EU FP6 funding (contract no LSHM-CT-2003-503041). SLD is funded by the BHF. MI, IB, and PD are funded by the Wellcome Trust. Some computation was done on the Vital-IT system at the Swiss Institute of Bioinformatics. The Lausanne and GEMS study were sponsored in part by GlaxoSmithKline. The support of Allen Roses, Lefkos Middleton, and Paul Matthews is greatly appreciated.

References

1 Baigent C, Keech A, Kearney PM, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *Lancet* 2005; **366:** 1267–78.

2 Rader DJ, Cohen J, Hobbs HH. Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. *J Clin Invest* 2003; **111:** 1795–803.

3 Heller DA, de Faire U, Pedersen NL, Dahlen G, McClearn GE. Genetic and environmental influences on serum lipid levels in twins. *N Engl J Med* 1993; **328:** 1150–56.

4 Topol EJ, Smith J, Plow EF, Wang QK. Genetic susceptibility to myocardial infarction and coronary artery disease. *Hum Mol Genet* 2006; **15** (spec 2)**:** R117–23.

5 Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6:** 95–108.

6 Kruglyak L. Power tools for human genetics. *Nat Genet* 2005; **37:** 1299–300.

7 Day N, Oakes S, Luben R, et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer* 1999; **80** (suppl 1)**:** 95–103.

8 Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 1972; **18:** 499–502.

9 Anon. Genetic information from the British 1958 birth cohort. DNA collection. http://www.b58cgene.sgul.ac.uk/collection.php (accessed Jan 28, 2008).

10 The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447:** 661–78.

11 Marques-Vidal P, Pecoud A, Hayoz D, et al. Prevalence and characteristics of vitamin or dietary supplement users in Lausanne, Switzerland: the CoLaus study. *Eur J Clin Nutr* published online Oct 17, 2007; DOI:10.1038/sj.ejcn.1602932.

12 Stirnadel H, Lin X, Ling H, et al. Genetic and phenotypic architecture of metabolic syndrome-associated components in dyslipidemic and normolipidemic subjects: The GEMS Study. *Atherosclerosis* published online Sept 20, 2007; DOI:10.1016/j.atherosclerosis.2007.07.038.

13 Loos RJ, Franks PW, Francis RW, et al. TCF7L2 polymorphisms modulate proinsulin levels and β-cell function in a British Europid population. *Diabetes* 2007; **56:** 1943–47.

14 Inouye M, Kumanduri V. WTSI high-throughput genotyping quality assessment and control. http://www.sanger.ac.uk/Teams/Team67/wtsi_qc.pdf (accessed Jan 14, 2007).

15 van Heel DA, Franke L, Hunt KA, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007; **39:** 827–29.

16 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55:** 997–1004.

17 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21:** 263–65.

18 Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002; **296:** 2225–29.

19 Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39:** 906–13.

20 Cambien F, Tiret L. Genetics of cardiovascular diseases: from single mutations to the whole genome. *Circulation* 2007; **116:** 1714–24.

21 Bennet AM, Di Angelantonio E, Ye Z, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* 2007; **298:** 1300–11.

22 Anon. Whole genome scan for type 2 diabetes in a Scandinavian cohort. http://www.broad.mit.edu/diabetes/scandinavs/index.html (accessed Jan 28, 2008).

23 Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. *Nature* 2007; **447:** 655–60.

24 Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007; **357:** 443–53.

25 Hingorani A, Humphries S. Nature's randomised trials. *Lancet* 2005; **366:** 1906–08.

26 Mahley RW. Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science* 1988; **240:** 622–30.

27 Horton JD, Cohen JC, Hobbs HH. Molecular biology of PCSK9: its role in LDL metabolism. *Trends Biochem Sci* 2007; **32:** 71–77.

28 Kotowski IK, Pertsemlidis A, Luke A, et al. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am J Hum Genet* 2006; **78:** 410–22.

29 Humphries SE, Whittall RA, Hubbart CS, et al. Genetic causes of familial hypercholesterolaemia in patients in the UK: relation to plasma lipid levels and coronary heart disease risk. *J Med Genet* 2006; **43:** 943–49.