Appendix 5: Details of statistical methods in the CRP CHD Genetics Collaboration (CCGC) [posted as supplied by author]

Statistical methods: All hypothesis tests were conducted using two-sided P-values and uncertainties about the point estimates were expressed using 95% confidence intervals (CIs). Studies contributing 50 or fewer participants, or 10 or fewer outcomes to any particular analysis were excluded. Primary analyses were conducted using an additive model for the effect of a pre-specified risk allele for each single nucleotide polymorphism (SNP), coded 0,1 or 2. Associations between SNPs and log_e CRP levels and other continuous markers were assessed by calculating the standardized differences (ie, standard deviation [SD] scale) in levels of each marker per additional copy of the risk allele, adjusted for ethnicity in a random effects meta-analysis model, allowing for heterogeneity across studies.[1] Details of the statistical methods used for the meta-analysis of exposure-outcome associations[2–4] and correction for potential bias due to regression dilution and within-person variability,[5] have been described previously.

SNPs and Haplotypes: Five SNPs (rs3093077, rs1205, rs1130864, rs1800947 and rs2794521) identified in previous work[6] as representing the variation in the CRP gene in European descent populations, were pre-specified for use in the current analyses. Using genetic data available on up to 37 further SNPs, we confirmed the validity of these tagging SNPs and generated study-specific haplotypes using fastPHASE software.[7] We found that the genotypes of 99% of individuals of European descent could be determined uniquely using 5 candidate haplotypes, and that the inclusion of further SNPs was only informative where data were not available on the tagging SNPs. Furthermore, only 4 SNPs were required to generate the haplotypes, as the 5th could be inferred by mutual exclusion. **Table A** in **appendix 1** lists the relevant haplotypes and haplotype frequencies. Where studies measured SNPs other than the 5 pre-specified tagging SNPs, we were able to use information on linkage disequilibrium (LD) from contributing studies, supplemented by SeattleSNP[8] and HapMap,[9] to assign participants to equivalent haplotype groups using different tagging SNPs. The LD threshold (r^2) for each pair of SNPs was defined, *a priori*, as 1 (ie, only tagging SNPs in complete LD with the principal SNPs were used to generate haplotypes). Sensitivity analyses using SNPs in lower LD did not enable the inclusion of any additional haplotype data. Only directly genotyped SNPs were used for both the haplotype and SNP-based analyses. Due to missing data in certain individuals on the rs3093077 SNP (or relevant tagging SNPs) it was not possible to discriminate between haplotypes 2 and 3. Therefore these individuals were allocated to haplotype 6. Similarly, individuals without the rs1800947 SNP were placed in haplotype 7 (as it was not possible to allocate these individuals to haplotypes 4 and 5, **Table A** in **appendix 1**). Individuals were excluded from haplotype analyses if insufficient SNP data were available to determine haplotypes, accounting for approximately 1% of the overall population. Due to differences in SNP LD, individuals of non-European descent, comprising approximately 9% of the available data (15,285 participants) were also excluded from haplotype analyses.

Haplotype associations with log$_e$ CRP levels and CHD risk were conducted within each study based on a model assuming additive effects across haplotypes, with two copies of haplotype 1 defined as the reference group. The study-specific estimates of the additive haplotype effects were then combined across studies by multivariate random effects meta-analysis to account for both the within and between study correlations in the haplotype effects while obtaining the pooled estimates and their uncertainty.[10] This model assumes that the effect of each haplotype is additive over haplotype 1. For example, based on our assumptions, individuals carrying 1 copy of haplotype 2 and 1 copy of haplotype 4 would have a mean log$_e$ CRP level equal to the sum of the haplotype 2 and haplotype 4 effects, relative to haplotype 1.

Analyses of incident and prevalent outcomes in prospective cohort studies: Where studies provided time-to-event data for incident cases and baseline data on prevalent CHD, we conducted a two sets of analyses: (i) Cox regression for time-to-incident CHD among the initially disease-free participants, and (ii) logistic regression of prevalent outcomes in the cohort with those who were disease-free at baseline within that study taken as controls. The two estimates obtained from the prospective study were then combined using inverse-variance weights, prior to combining with estimates from other studies in the overall meta-analyses. These methods enabled us to maximize the amount of data available for analyses of SNP-CHD associations and haplotype-CHD associations. Sensitivity analyses excluding prevalent cases or using only logistic regression for all cases were conducted to test these assumptions, and no substantial differences were seen. Analyses of SNP-CRP, haplotype-CRP and CRP-CHD associations were limited to participants without known cardiovascular disease at time of CRP measurements.

Mendelian randomization analyses: Using a Bayesian framework with vague independent $\mathcal{N}(0, 10^2)$ priors on average phenotype level and regression parameters, and vague independent $\mathcal{U}(0, 20)$ priors on standard deviation parameters in normal distributions, we generated a causal estimate for circulating CRP on CHD, using a SNP-based method and a haplotype-based method for predicting levels of CRP.[11] Within each study, the participants were subdivided into groups based on genotype or haplotype. Genotypes were defined by the hierarchical addition of up to 4 SNPs: rs3903077, rs1205, rs1130864 and rs1800947. Genotype groups were also subdivided by ethnicity (European descent [ED], African descent [AD] or Asian descent [OD]).

For each subgroup *j*, we estimated the mean level of phenotype ξ_j assuming that, for each individual *i* in subgroup *j*, the measured values of phenotype x_{ij} came from a normal distribution with mean ξ_j and variance σ^2, assumed to be common across subgroups. In the SNP-based analysis, the mean level of phenotype ξ_j for people with g_{jk} variant allele copies for each SNP $k(1 \leq k \leq K)$, was estimated using an additive model, with α_k representing the change in phenotype per allele change in SNP k:

$$\xi_j = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{jk} \tag{1}$$

In the haplotype-based analysis, the mean level of phenotype ξ_j for people with haplotypes h_{1j} and h_{2j}, was similarly estimated using additive effects for each haplotype.

The probability of an event in subgroup j, was estimated assuming a binomial model for the number of events n_j from total number at risk N_j. Using a logistic model, and assuming a linear relationship between the log-odds of an event and mean levels of phenotype ξ_j, the coefficient β_1 for an increase in log-odds of event per unit increase in phenotype was estimated as follows:

$$x_{ij} \sim \mathcal{N}(\xi_j, \sigma^2)$$

$$n_j \sim \mathcal{B}(N_j, \pi_j)$$

$$\text{logit}(\pi_j) = \beta_0 + \beta_1 \xi_j \tag{2}$$

We now consider multiple studies, indexed by m. Where studies have not measured (or shared data on) circulating CRP, we used the combined estimates of genetic association α_{km} with CRP across studies that provided relevant data, assuming the change in phenotype per allele change in SNP is consistent across studies. We used a random effects meta-analysis model that imposed a normal distribution on study-level parameters α_{km} with mean $\mu_{\alpha k}$ and variance ψ_k^2 for each SNP $k(1 \leq k \leq K)$. The resulting predictive distributions for the gene - CRP association parameters formed an implicit prior for the studies without circulating CRP data:

$$\xi_{jm} = \alpha_{0m} + \sum_{k=1}^{K} \alpha_{km} g_{jkm}$$

$$\alpha_{km} \sim \mathcal{N}(\mu_{\alpha k}, \psi_k^2) \tag{3}$$

To obtain a single causal estimate of the CRP - CHD association across studies, we conducted both a fixed effect and random effects meta-analysis. In the fixed effect meta-analysis, the causal parameter, β_1 was the same for each study m:

$$x_{ijm} \sim \mathcal{N}(\xi_{jm}, \sigma_m^2)$$

$$n_{jm} \sim \mathcal{B}(N_{jm}, \pi_{jm})$$

$$\text{logit}(\pi_{jm}) = \beta_{0m} + \beta_1 \xi_{jm} \tag{4}$$

In the random-effects meta analysis, a normal distribution was imposed on the study-level causal effect parameters, with the mean of this distribution taken as the parameter of interest. Study-level estimates from AD and OD populations were combined only using a fixed effect model due to insufficient data. All Bayesian models were fitted using WinBUGS.[12] Convergence was assessed through visual examination of trace plots, by assessment of Monte Carlo error, and by running multiple chains from dispersed starting values. At least 60 000 iterations were used for all analyses, with the first 10 000 iterations discarded as "burn-in".

Strength of instruments: Instrument strength was assessed by F statistics in the regression of log_e(CRP) levels on the genetic instruments. The F statistic for the SNP-based model for all participants is 89.2 on (16, 98460) degrees of freedom and for the haplotype-based model for European Descent populations is 239.8 on (6, 94516) degrees of freedom.

Censoring of outcomes: For participants who had multiple events (e.g. two CHD events at separate time points, or a CHD event followed by another type of event such as a stroke or death from cancer), analyses in the CCGC focused on first events.[13] Thus, in an analysis of CHD events, participants were followed until their first event, or censored at the time of other non-fatal cardiovascular events, such as stroke, or death from other causes. Individuals were not censored at the time of cardiovascular investigations or interventions, such as angiography or coronary bypass operations, or at the diagnosis of angina. The rationale for this was that major cardiovascular events, such as first non-fatal MI or stroke, may disrupt the association between baseline risk factors and subsequent disease risk. The incidence of angina and coronary interventions was, however, not recorded reliably enough in sufficient studies to consider censoring for them. The potential biases that arise through these decisions on censoring were addressed through sensitivity analyses and by implementing alternative censoring criteria. In general, such changes had only minimal effects.

Inclusion of tabular data: Some of the studies listed in **Table B** in **appendix 1** elected to share tabular data on CRP SNPs and CHD risk. These data were included in the SNP-based analyses and incorporated via a univariate meta-analyses of risk ratios for CHD. Similarly, these estimates were incorporated into the Bayesian estimates of the causal role of CRP in CHD (using genetic data, as outlined above). Due to the unavailability of plasma CRP measures in the these studies, the mean CRP levels in each genotype group (combined from the remaining studies) were used as estimates, as described above. Sensitivity analyses excluding studies that shared tabular data did not alter the findings.

Important conversions: A unit higher log_e CRP corresponds approximately to a 3-fold higher CRP: the standard deviation of log_e CRP level was 1.057, and $e^{1.057} = 2.878$. Conversions were always done to at least three decimal places for accuracy.

Long-term associations of SNPs with circulating CRP: To assess whether CRP genetic variants consistently correlate with circulating CRP levels over time, we used repeat measures of CRP from participating studies to assess the evidence of SNP*time interactions. Thus for each study with repeat CRP measurements, we regressed repeat measurements of log_e CRP, on SNP, time, SNP*time and ethnicity, and combined the study-specific estimates using random effects meta-analysis. **Figure D** in **appendix 3** shows the SNP-CRP association plotted over time using the combined regression coefficients: the relationships between CRP SNPs and circulating log_e CRP remain fairly constant over time.

Sensitivity analyses: To investigate the reliability of the results, a number of sensitivity analyses were undertaken. We repeated the primary analyses excluding studies and SNPs out of Hardy Weinberg Equilibrium, and with an F-statistic less than 10. In order to verify that our choice of an additive model for SNP effects was appropriate, we conducted a model free analysis (ie, 2 degrees of freedom test) of SNP - CRP and SNP - CHD effects (**Figure L** in **appendix 3**). The ratio (referred to as λ) of the effect contrast between the heterozygotes and homozygotes (versus common homozygotes), can range from 0, suggesting a recessive model, to 1, suggesting a dominant model. Values of 0.5 correspond to a co-dominant or additive model.[14] For these analyses, λ was estimated from the summary estimates for the SNP - CRP associations, on the assumption that these would better represent the appropriate model for Mendelian randomization analyses. As shown in

Figure L in **appendix 3**, the additive model was appropriate for all four of the principal SNPs. All analyses were conducted using random and fixed effects analyses (the latter presented in **Figures C, G, I** and **K** in **appendix 3**). Finally, subgroup analyses were conducted for each step of the analyses, including by different age groups, by sex, ethnicity and geographical region, genotyping and assay methods, and excluding participants on cardiovascular medications at baseline. Details on methods used for subgroup analyses are presented separately.[15]

References

[1] J. Higgins, S. Thompson, and D. Spiegelhalter, "A re-evaluation of random-effects meta-analysis," *J Roy Stat Soc A Stat Soc*, 2009;172:137-59.

[2] Emerging Risk Factors Collaboration, "C-reactive protein ceoncentration and risk of coronary heart disease, stroke and mortality: an individual participant meta-analysis," *Lancet*, 2010; 375:132-40.

[3] Emerging Risk Factors Collaboration, "Major lipids, apolipoproteins and risk of vascular disease," *JAMA*, 2009;302(18):1993-2000.

[4] Emerging Risk Factors Collaboration, "Lipoprotein(a) concentration and the risk of coronary heart disease, stroke and nonvascular mortality," *JAMA*, 2009;302(4):412-23.

[5] Fibrinogen Studies Collaboration, "Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies.," *Stat Med*, 2008;28(7):1067-1092.

[6] C. Verzilli, T. Shah, J. Casas, J. Chapman, M. Sandhu, S. Debenham, M. Boekholdt, K. Khaw, N. Wareham, R. Judson, *et al.*, "Bayesian meta-analysis of genetic association studies with different sets of markers," *Am J Hum Genet*, 2008;82(4):859-72.

[7] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *Am J Hum Genet*, 2006;79:629-44.

[8] SeattleSNPs, "SeattleSNPs NHLBI Program for Genomic Applications." URL: http://pga.gs.washington.edu, last accessed November 2009.

[9] The International HapMap Consortium, "The international HapMap project," *Nature*, 2003;426:789-96.

[10] I. White, "Multivariate random-effects meta-analysis," *Stata J*, 2009;9(1):40-56.

[11] S. Burgess, S. G. Thompson, CHD CRP Genetics Collaboration, "Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables," *Stat Med*, 2010;29(12):1298-311.

[12] D. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility," *Stat Comput*, 2000;10(4):325-37.

[13] CRP CHD Genetics Collaboration, "Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation," *Eur J Epidemiol*, 2008;23(8):531-40.

[14] C. Minelli, J. Thompson, K. Abrams, A. Thakkinstian, and J. Attia, "The choice of a genetic model in the meta-analysis of molecular association studies," *Int J Epidemiol*, 2005;34(6):1319.

[15] Emerging Risk Factors Collaboration, "Statistical methods for the analysis of individual participant data from multiple epidemiological studies," *Int J Epidemiol*, 2010;39(5):1345-59.