Efficient Calculation for Multi-SNP Genetic Risk Scores

Toby Johnson `<Toby.x.Johnson@gsk.com>`, GlaxoSmithKline, Stevenage, UK

presented at the American Society of Human Genetics Annual Meeting, San Francisco, November 6–10, 2012

1 Motivation

Genetic risk scores based on genotypes at multiple single nucleotide polymorphisms (SNPs) have several applications in association studies for complex human phenotypes. However, for many human diseases and traits of clinical importance, identifying genetic associations has required large sample sizes, so single SNP associations are typically studied by meta-analysis of summary results obtained from multiple genotype-phenotype datasets held at different institutions. In such settings, testing association between a multi-SNP genetic risk score and a phenotype is prone to organisational difficulties and potential for analytic error.

One application of interest is estimating the causal effect of a chosen biomarker on a disease outcome, invoking a Mendelian randomisation argument [LHS$^+$08]. An investigator may have access to information sufficient to parameterise a genetic risk score for the biomarker, but may not have direct access to genotype data sufficient to test association between the risk score and the disease. In this application, it is desirable to make efficient use of single SNP meta-analysis association statistics, calculated by research consortia using datasets with very large total sample sizes of disease case and control subjects [e.g. SKK$^+$11, MVT$^+$12].

The key quantities of interest are the effect size estimate for association between the risk score and the disease (here denoted $\hat{\alpha}$), and its standard error (SE). These are typically sufficient to calculate other quantities of interest, such as the association P-value, (pseudo-)variance explained, and the ratio estimate for the causal effect of a biomarker on the disease.

2 Definitions

For an additive multi-SNP risk score depending on m chosen SNPs, the value of the risk score for the i-th subject is

$$r_i = \sum_{j=1}^{m} w_j x_{ij} \quad . \tag{1}$$

Here, x_{ij} is the dose of the coded allele at the j-th SNP in the i-th subject, and w_j is a chosen coefficient or weight for the j-th SNP. The choice of the SNPs and the vector of coefficients $\boldsymbol{w}$ together parameterise the score and are assumed known.

Assume we wish to assess association between the risk score (1) and a chosen phenotype in a chosen dataset, using a regression model where the likelihood [or the partial likelihood for a Cox proportional hazards model] of the observed phenotype data depends on explanatory variables only through a linear predictor $\eta_i = r_i\alpha + \cdots$.

The key quantity of interest is $\hat{\alpha}$, an estimate for α, the coefficient for the risk score in the linear predictor.

3 Results

3.1 "Summary statistic" method

Define

$$\tilde{\alpha} := \frac{\sum_{j=1}^{m} w_j \hat{\beta}_j s_j^{-2}}{\sum_{j=1}^{m} w_j^2 s_j^{-2}} \qquad \text{with} \qquad \mathrm{SE}(\tilde{\alpha}) = \sqrt{\frac{1}{\sum_{j=1}^{m} w_j^2 s_j^{-2}}} \quad , \tag{2}$$

where $\hat{\beta}_j$ is the effect size estimate when the phenotype is regressed onto x_{ij} in a single SNP analysis in the chosen dataset, and $s_j = \mathrm{SE}\left(\hat{\beta}_j\right)$ is the corresponding SE.

3.2 Main result

When all SNP genotypes used in the risk score are uncorrelated, then

$$\tilde{\alpha} \simeq \hat{\alpha} \tag{3}$$

Trivially, under the null hypothesis that the $\hat{\beta}_j$ are independently normally distributed with means zero and variances s_j^2, the quantity $\tilde{\alpha}$ is normally distributed with mean zero and variance $\mathrm{SE}(\tilde{\alpha})^2$.

3.3 Goodness of fit test

When all SNP genotypes used in the risk score are uncorrelated, then

$$X_m^2 := \sum_{i=1}^{m} \hat{\beta}_j^2 s_j^{-2} \qquad \text{and} \qquad X_{\mathrm{rs}}^2 := \left(\frac{\tilde{\alpha}}{\mathrm{SE}(\tilde{\alpha})}\right)^2 \tag{4}$$

are $\chi^2_{(m)}$ and $\chi^2_{(1)}$ distributed test statistics for association between the phenotype and all m SNPs under an unconstrained m d.f. model, and for the nested 1 d.f. risk score model, respectively. Then,

$$Q_{\mathrm{rs}} := X_m^2 - X_{\mathrm{rs}}^2 \tag{5}$$

is $\chi^2_{(m-1)}$ distributed, under the null hypothesis that all m SNPs are associated with the phenotype with true effect sizes that are proportional to the coefficients $\boldsymbol{w}$ used to parameterise the risk score. This null hypothesis expresses a critical assumption required for a Mendelian randomisation argument, namely that all the genetic instruments must affect disease risk *only* through their effects on the biomarker of interest, and must not have other "pleiotropic" effects on disease risk (e.g. via other biomarkers).

3.4 Proof of main result

Here I write the proof for (3) only for the simplest case, where the chosen phenotype is a continuous trait (z_i in the i-th subject), and where there are no covariates. Then, for the j-th SNP, the regression coefficient and SE are

$$\hat{\beta}_j = \frac{\boldsymbol{z}'\boldsymbol{x}_j}{\boldsymbol{x}_j{}'\boldsymbol{x}_j} \qquad s_j \simeq \sqrt{\frac{\boldsymbol{z}'\boldsymbol{z}}{n\,\boldsymbol{x}_j{}'\boldsymbol{x}_j}} \tag{6}$$

where $\boldsymbol{z}$ is the centered $n \times 1$ vector of subject-specific trait values and $\boldsymbol{x}_j$ is the centered $n \times 1$ vector of coded allele dosages. The approximation for s_j in (6) assumes n is large and that a small fraction of the trait variance is explained. For the risk score the regression coefficient is

$$\hat{\alpha} = \frac{\boldsymbol{z}'\boldsymbol{r}}{\boldsymbol{r}'\boldsymbol{r}} \tag{7}$$

where $\boldsymbol{r}$ is the centered $n \times 1$ vector of subject-specific multi-SNP risk score values. The required result

$$\hat{\alpha} = \frac{\boldsymbol{z}'\boldsymbol{r}}{\boldsymbol{r}'\boldsymbol{r}} \simeq \frac{\sum_{j=1}^{m} w_j \boldsymbol{z}'\boldsymbol{x}_j}{\sum_{j=1}^{m} w_j^2 \boldsymbol{x}_j{}'\boldsymbol{x}_j} \simeq \frac{\sum_{j=1}^{m} w_j \hat{\beta}_j s_j^{-2}\, n\,(\boldsymbol{z}'\boldsymbol{z})^{-1}}{\sum_{j=1}^{m} w_j^2 s_j^{-2}\, n\,(\boldsymbol{z}'\boldsymbol{z})^{-1}} = \tilde{\alpha} \tag{8}$$

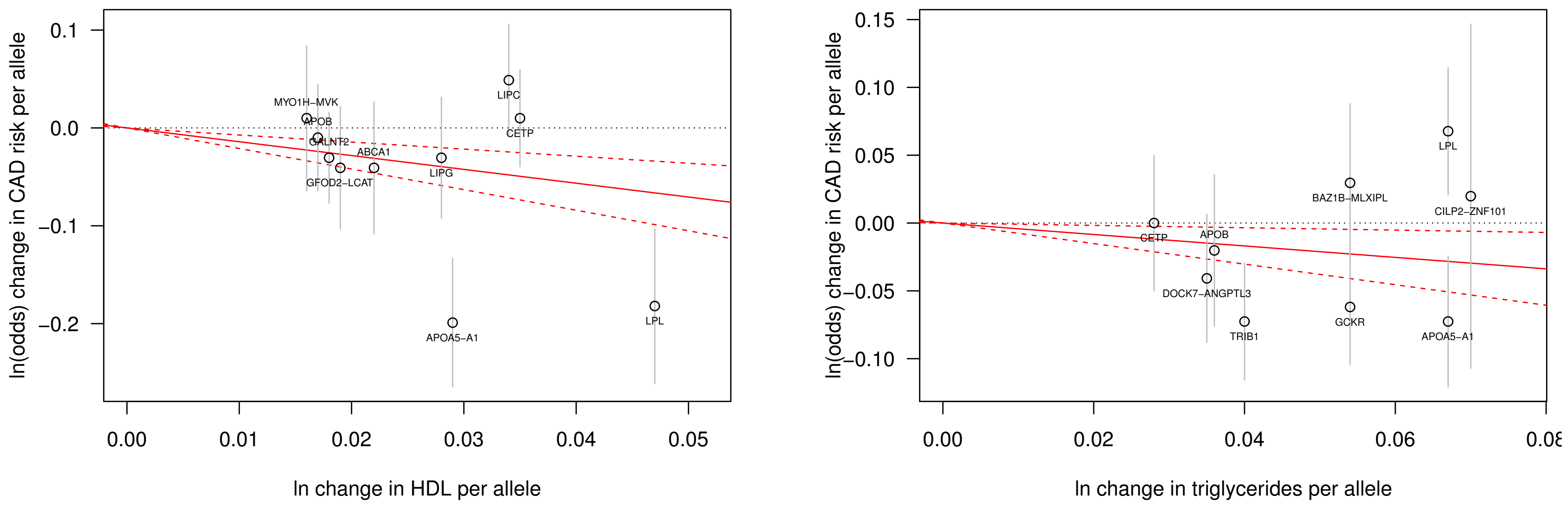
has two necessary conditions: (i) that $\boldsymbol{r} = \sum_{j=1}^{m} w_j \boldsymbol{x}_j$, which is true because the [centered] risk score (1) is a linear combination of the [centered] coded allele dosages; and (ii) that $\boldsymbol{x}_j{}'\boldsymbol{x}_k \simeq 0$ for all $j \neq k$, that is that the centered vectors of coded allele dosages are orthogonal, which is true for uncorrelated SNP genotypes. □

I wrote a more general but less rigorous derivation for (3) in [DHT$^+$12], and I have performed extensive numerical verification of (3) using Monte Carlo subsamples of several real datasets.

4 Illustrative application

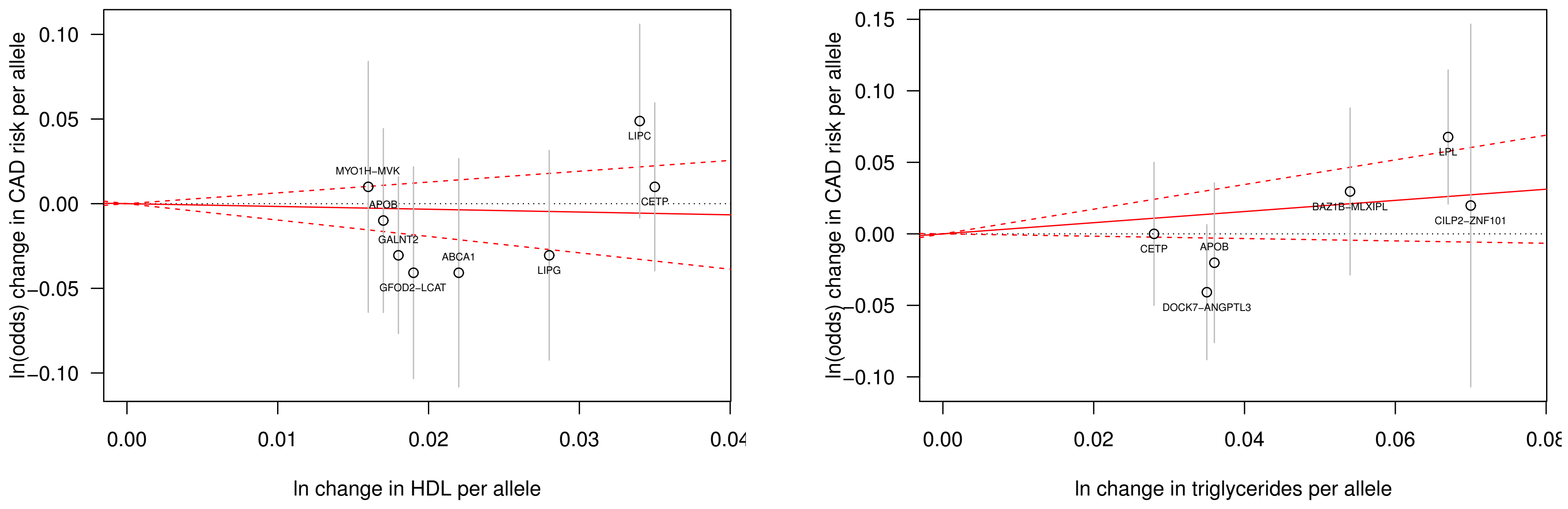
Previously, [WRS$^+$10] reported effect size estimates for association between prevalent coronary artery disease (CAD) and genotypes at 29 SNPs associated with serum lipid biomarkers, combining results from nine studies totalling 9 633 CAD cases and 38 684 controls. Effect size estimates for association between the biomarkers (low density lipoprotein cholesterol (LDL), high density lipoprotein cholesterol (HDL), and triglycerides) and the genotypes were also reported, combining results from eight studies totalling 17 723 subjects (partly overlapping the CAD case and control samples). However, [WRS$^+$10] did not report any results for multi-SNP risk scores.

Figure 1. Estimated effects on CAD risk are plotted against estimated effects on serum lipid biomarkers, for ten SNPs associated with HDL (left panel) and for nine SNPs associated with triglycerides (right panel). Vertical grey lines show 95% confidence interval (CI) for each individual SNP. Estimates of casual effect of each biomarker on CAD risk, by applying (2) using all SNPs in each panel, are represented by red solid lines with gradient $\tilde{\alpha}$, with red dashed lines showing the 95% CI.



Using all SNPs, multi-SNP risk score analyses identify weak but statistically significant protective causal effects of HDL and triglycerides on CAD risk (0.87 odds (95% CI 0.82–0.93) per 10% increase in HDL, $P = 6.1 \times 10^{-5}$; 0.96 odds (95% CI 0.93–0.99) per 10% increase in triglycerides, $P = 0.014$). However, applying (5) detects strong evidence of heterogeneity of effects on CAD risk relative to the estimated effects on either biomarker ($Q_{\mathrm{rs}} = 48.39$ on 9 d.f., $P = 2.2 \times 10^{-7}$ for HDL; $Q_{\mathrm{rs}} = 34.12$ on 8 d.f., $P = 3.9 \times 10^{-5}$ for triglycerides). Hence the 1 d.f. risk score models do not fit these data, and the assumptions required for a Mendelian randomisation argument must be seriously questioned.

Figure 2. Stepwise removal of SNPs from the risk score, minimising Q_{rs} at each step until there was no significant heterogeneity (at $P \leq 0.05$), removed SNPs at the *APOA5-A1* and *LPL* loci for HDL, and removed SNPs at the *GCKR*, *TRIB1* and *APOA5-A1* loci for triglycerides. Estimated effects on CAD risk are plotted against estimated effect on serum lipid biomarkers for the remaining SNPs. Corresponding estimates of casual effect of biomarker on CAD risk are represented by red lines as before.



After removing SNPs that likely violate assumptions required for a Mendelian randomisation argument, multi-SNP risk score analyses suggest no causal effects of either HDL or triglycerides on CAD risk, with narrow 95% CIs (0.91–1.06 odds for CAD per 10% HDL increase, $P = 0.69$; 0.91–1.01 odds for CAD per 10% triglyceride decrease, $P = 0.11$).

By constructing a genetic risk score for HDL, excluding "pleiotropic" SNPs significantly associated (at $P \leq 0.01$) with either LDL or triglycerides, [VPOM$^+$12] used the method described here to support a related conclusion, that HDL has no causal effect on myocardial infarction risk. An advantage of the the goodness of fit approach described here is that SNPs with "pleiotropic" effects on disease risk can be excluded from the risk score, even when suitable markers for the "pleiotropic" effects are not available.

References

[DHT$^+$12] Z. Dastani et al. (2012) Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genetics* **8**:e1002607. doi:10.1371/journal.pgen.1002607.

[LHS$^+$08] D. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson and G. Davey Smith (2008) Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**(8):1133–1163. doi:10.1002/sim.3034.

[MVT$^+$12] A. P. Morris et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**(9):981–990. doi:10.1038/ng.2383.

[SKK$^+$11] H. Schunkert et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**(4):333–338. doi:10.1038/ng.784.

[VPOM$^+$12] B. F. Voight et al. (2012) Plasma HDL cholesterol and risk of myocardial infarction: A Mendelian randomisation study. *The Lancet* **380**(9841):572–580. doi:10.1016/S0140-6736(12)60312-2.

[WRS$^+$10] D. M. Waterworth et al. (2010) Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arteriosclerosis, Thrombosis, and Vascular Biology* **30**:2264–2276. doi:10.1161/ATVBAHA.109.201020.

Methods described here are implemented in my `gtx` package for the R statistical programming language, available at `http://cran.r-project.org/web/packages/gtx`