Genetics

Value of genetic profiling for the prediction of coronary heart disease

Jeroen B. van der Net, MD, PhD, [a,b] A. Cecile J.W. Janssens, PhD, [a] Eric J.G. Sijbrands, MD, PhD, [b] and Ewout W. Steyerberg, PhD [a] *Rotterdam, The Netherlands*

Background   Advances in high-throughput genomics facilitate the identification of novel genetic susceptibility variants for coronary heart disease (CHD). This may improve CHD risk prediction. The aim of the present simulation study was to investigate to what degree CHD risk can be predicted by testing multiple genetic variants (genetic profiling).

Methods   We simulated genetic profiles for a population of 100,000 individuals with a 10-year CHD incidence of 10%. For each combination of model parameters (number of variants, genotype frequency and odds ratio [OR]), we calculated the area under the receiver operating characteristic curve (AUC) to indicate the discrimination between individuals who will and will not develop CHD.

Results   The AUC of genetic profiles could rise to 0.90 when 100 hypothetical variants with ORs of 1.5 and genotype frequencies of 50% were simulated. The AUC of a genetic profile consisting of 10 established variants, with ORs ranging from 1.13 to 1.42, was 0.59. When 2, 5, and 10 times as many identical variants would be identified, the AUCs were 0.63, 0.69, and 0.76.

Conclusion   To obtain AUCs similar to those of conventional CHD risk predictors, a considerable number of additional common genetic variants need to be identified with preferably strong effects. (Am Heart J 2009;158:105-10.)

The genetics of coronary heart disease (CHD) has received great interest over the last decades with the latest developments in high-throughput genomics and genome-wide association studies.[1-4] Not only will these advances lead to better understanding of the etiology of CHD, it is also expected that genetic profiling, that is, the simultaneous testing of multiple genetic variants, can eventually be used to predict CHD risk in individuals. This may lead to personalized medicine in which preventive and therapeutic interventions will be targeted at genetic profiles rather than at conventional risk factors. A potential advantage of genetic profiling above conventional risk factors in the prediction of CHD risk is that genetic information is already present at birth. Therefore, genetic profiles could theoretically predict CHD risk before conventional risk factors become apparent. The expectations are high,[5-7] but the question remains whether genetic profiling will indeed become useful for the prevention and treatment of CHD.

A recent study investigated the combined influence on CHD risk of a genetic profile consisting of 10 genetic variants that were all significantly associated with CHD in published meta-analyses.[5] In a simulation study, the authors determined the expected frequency of individuals with different numbers of risk genotypes and their effect on CHD risk. Although this study showed that individuals with a high number of risk genotypes had a very high CHD risk compared with individuals with a low number of risk genotypes, the authors did not investigate whether this genetic profile discriminated between individuals who will develop CHD and those who will not. Discriminative accuracy is the first requirement for a valuable predictive test and can be expressed as the area under the receiver operating characteristic curve (AUC).[8] Previous studies that investigated the discriminative accuracy for prediction of CHD generally found small discriminative accuracies of genetic profiles (AUC in the order of 0.60, denoting little discrimination).[9,10] Whereas these previous studies only considered genetic profiles consisting of a limited number of genetic variants, most likely, a much larger number of genetic variants with small effects is involved in CHD.[11] Inclusion of more genetic variants in genetic profiles may further improve their discriminative accuracy.

The aim of the present study was to investigate what we can expect from genetic profiling in the prediction of CHD with the currently known genetic variants and when more genetic variants associated with CHD will be identified in

the future. We compared the discriminative accuracy of genetic profiles with that of CHD prediction models based on conventional risk factors. Because limited empirical data are available of large-scale genetic profiling, simulation studies were required for the present analyses.

Methods

Modeling strategy

We used a modeling procedure that created a dataset with information on genetic profiles and disease status for a population of 100,000 individuals. We simulated genetic profiles that consisted of multiple genetic variants, which can be polymorphisms or haplotypes.

The modeling procedure has been published in detail elsewhere[11] but is briefly as follows: For the construction of the genetic profiles, we assumed (1) that each genetic variant only had a risk variant and a wild-type variant, therefore, interpretable as a dominant or recessive mode of inheritance, (2) that genetic variants inherited independently, that is, no linkage disequilibrium existed between the genetic variants, and (3) that the combined effect of the genetic variants on disease risk followed a multiplicative risk model of independent genetic effects, that is, no statistical interaction terms were included in the model.[12,13]

The modeling was done in 3 steps: (1) modeling the genetic profiles of all individuals, (2) calculating the cumulative 10-year CHD risks associated with the genetic profiles, and (3) modeling the CHD status of all individuals. We constructed the genetic profiles by randomly assigning the genotypes of each genetic variant to all individuals in the population, so that the genotype distributions are in line with the specified genotype frequencies. The CHD risks associated with the genetic profiles were calculated using Bayes theorem. Bayes theorem states that the posterior odds of CHD for each individual are obtained by multiplying the prior odds by the likelihood ratio (LR) of their test result. The prior odds are calculated from the baseline population CHD risk (p) using the formula $p/(1 - p)$. The LR of the test result refers to the LR of the genetic profile,[14] which, under the assumption of independent genetic effects, was obtained by multiplying the LRs of all individual genotypes that constitute the profile. The LRs of the genotypes of the single genetic variants were calculated from a 2-by-2 contingency table presenting genotype by disease status. This table is constructed from the genotype frequency, the population CHD risk, and the strength of association between the risk genotype of a genetic variant and CHD risk, quantified by the odds ratio (OR). These are all specified as parameters that vary between different simulation scenarios (see below). The posterior odds are converted into CHD risks using the formula odds/(1 + odds). Finally, to model CHD status, we used a simulation procedure that assigns disease status to individuals based on the CHD risks associated with their genetic profiles. This procedure ensures that for individuals with the same disease risk, the percentage of individuals who will develop CHD equals that CHD risk, when the subgroup of individuals with that CHD risk would have been sufficiently large.[11]

Discriminative accuracy

The discriminative accuracy is the extent to which test results can discriminate between individuals who will develop CHD and those who will not.[8,15] The AUC is commonly used to quantify the discriminative accuracy of a prediction model.[16] The AUC is the probability that the test correctly identifies the diseased individual from a pair of whom one is affected and one is unaffected and ranges from 0.5 (total lack of discrimination) to 1.0 (perfect discrimination).[17] The AUC was obtained as the c-statistic by the R function *somers*2, which is available in the Hmisc library of R software.[18] Simulation studies can simulate any number of genes with any strength of effect. Yet, the contribution of genetic factors in the development of CHD is bounded, as indicated by the heritability. In each scenario, we calculated the proportion of explained variance (R^2) as a proxy measure of the heritability, that is, the total contribution of genetic factors in the specific scenario.[11] All simulations were repeated 10 times to obtain robust estimates of the AUC and proportion of explained variance; all results are presented as averages of the repeated simulations. Analyses were performed using R software (version 2.5.1).[18]

Simulation scenarios

We considered 3 different scenarios, in each of which we simulated genetic profiles and CHD status for 100,000 individuals, and a 10-year cumulative CHD risk of 10%.

First, to gain a first impression of the discriminative accuracy of genetic profiles, we simulated genetic profiles consisting of 50 and 100 genetic variants that all had the same OR and genotype frequency. For each genetic profile, we obtained the distribution of the number of risk genotypes in the population. Subsequently, we investigated the AUCs of the genetic profiles. In separate simulations, we considered different combinations of ORs (1.1, 1.2, and 1.5) and genotype frequencies (10%, 30%, and 50%).

Second, we considered genetic profiles defined by 10 established genetic variants of which the combined effect on CHD risk was recently studied.[5] We constructed genetic profiles consisting of genetic variants that had ORs and genotype frequencies identical to the 10 established genetic variants. We calculated the AUC of simultaneously testing these 10 variants, and additionally, we estimated the expected AUC when 2, 5, and 10 times as many genetic variants with the same distribution of ORs and genotype frequencies would be known, that is, genetic profiles consisting of 20, 50, and 100 genetic variants in total.

Finally, we investigated conditions in which genetic profiling yields a similar discriminative accuracy as conventional CHD risk prediction scores. We hereto compared the AUC of genetic profiles with that of the Framingham and PROCAM risk scores.[19,20] The AUCs of these scores were between 0.75 and 0.82 in the original studies,[20,21] but subsequent studies have reported AUCs of 0.61 to 0.68 in European populations.[22,23] For different genotype frequencies, we investigated the magnitude of the ORs of 1 to 100 additional genetic variants needed to obtain an AUC of 0.65, 0.70, or 0.75 in a panel already consisting of the 10 genetic variants described in the second scenario.

Results

Genetic profiling of 50 genetic variants with genotype frequencies of 10%, 30%, and 50% resulted in a mean

Figure 1
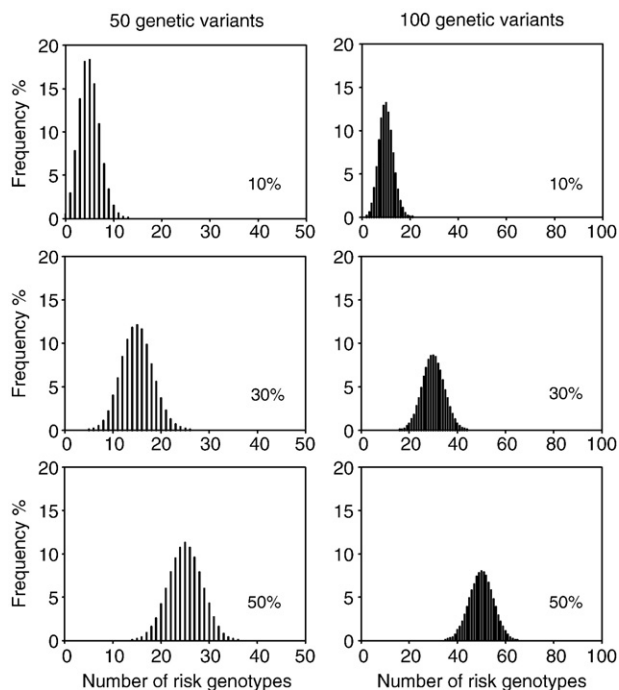
Distributions of the genetic profiles. Distributions of the genetic profiles in simulated populations as a function of the risk genotype frequency (upper panel 10%, middle panel 30%, and lower panel 50%). Genetic profiles were defined by 50 (left panel) or 100 genetic variants (right panel). All genes involved in the genetic profiles had the same ORs and risk genotype frequencies. The x-axis indicates the number of risk genotypes of 50 or 100 genetic variants in the genetic profiles. The y-axis indicates the percentage of individuals in the simulated population, for example, 18% of the population carries 4 risk genotypes when the genotype frequency is 10% and the genetic profile consists of 50 genetic variants.

number of risk genotypes in the simulated population of 5, 15, and 25, respectively (Figure 1). The observed range of the number of risk genotypes was 0 to 13 when the genotype frequency was 10% and 14 to 36 when the genotype frequency was 50%. These numbers were more or less doubled for genetic profiles consisting of 100 genetic variants. As expected, higher genotype frequencies, higher ORs, and larger numbers of genetic variants yielded a higher discriminative accuracy (Table I). The AUC of genetic profiles ranged from 0.56 when 50 genetic variants with ORs of 1.1 and genotype frequencies of 10% were simulated to 0.90 when 100 genetic variants with ORs of 1.5 and genotype frequencies of 50% were simulated (Table I). Similarly, the proportion of explained variance ranged from 0.5% to 33% (Table I).

Table I. Discriminative accuracy as a function of the OR and risk genotype frequency

OR	Genotype frequency*	AUC 50 genes	R^2	AUC 100 genes	R^2
1.1	10%	0.56	0.5%	0.59	1.0%
	30%	0.59	1.1%	0.63	2.1%
	50%	0.61	1.5%	0.64	2.6%
1.2	10%	0.62	1.9%	0.66	3.8%
	30%	0.68	4.7%	0.74	8.3%
	50%	0.69	4.7%	0.75	9.1%
1.5	10%	0.75	10%	0.82	19%
	30%	0.83	20%	0.89	32%
	50%	0.84	21%	0.90	33%

All genes involved in the genetic profiles have the same ORs and risk genotype frequencies. R^2, proportion of explained variance by genetic variants.
* Risk genotype frequency.

Table II. Established genetic variants from published meta-analyses

Gene, variant	Risk genotype	OR	Risk genotype frequency
ACE, insertion/deletion intron 16	D/D	1.22	28%
APOB, Q4154K	K+	1.32	27%
APOE, E2/E3/E4	E4+	1.42	23%
CETP, Taq1b	B1+	1.24	68%
ITGA2B, P1(A2)	A2+	1.13	27%
LPL, S447X	S/S	1.25	80%
MTHFR, C677T	T/T	1.14	12%
NOS3, E298D	D/D	1.31	11%
PON1, Q192R	R192+	1.15	21%
SERPINE1, 5G/4G	4G/4G	1.20	31%

Data obtained from Drenos et al.[5]

The discriminative accuracy was 0.59 for a genetic profile based on the 10 genetic variants that were significantly associated with CHD in published meta-analyses (Table II). When 2, 5, and 10 times as many genetic variants with the same distribution of ORs and genotype frequencies would be identified, the expected AUCs were 0.63, 0.69, and 0.76, respectively.

Table III shows the magnitude of the ORs and genotype frequencies of genetic variants that are needed in addition to the original genetic profile of 10 genetic variants to obtain AUCs similar to that of the conventional CHD risk prediction scores. For instance, an AUC of 0.75 could be obtained when 10 additional genetic variants would be identified that have an OR of at least 1.8, depending on the genotype frequency. For 20, 50, or 100 additional genetic variants, the minimal ORs were 1.5, 1.3, and 1.2, depending on the genotype frequency (Table III).

Discussion

An important factor in the potential success of genetic profiling in the prediction of CHD is the discovery of

Table III. Minimal ORs needed to obtain AUCs comparable with that of the Framingham and PROCAM risk scores in addition to the 10 genetic variants (AUC = 0.59) in Table II

Risk genotype frequency	No. of additional genetic variants	AUC 0.65	AUC 0.70	AUC 0.75
0.05	1	3.9	7.3	13.6
	5	2.1	2.8	3.8
	10	1.7	2.1	2.7
	20	1.5	1.8	2.2
	50	1.3	1.5	1.7
	100	1.2	1.3	1.5
0.10	1	2.8	4.8	7.5
	5	1.8	2.3	2.8
	10	1.5	1.9	2.2
	20	1.4	1.6	1.8
	50	1.2	1.4	1.5
	100	1.1	1.2	1.3
0.30	1	2.2	3.3	5.0
	5	1.5	1.8	2.2
	10	1.4	1.6	1.8
	20	1.3	1.4	1.5
	50	1.2	1.3	1.3
	100	1.1	1.1	1.2
0.50	1	2.2	3.5	6.7
	5	1.4	1.8	2.2
	10	1.3	1.5	1.8
	20	1.2	1.4	1.5
	50	1.1	1.2	1.3
	100	1.1	1.1	1.2

genetic variants that demonstrate consistent associations with CHD. We have shown that genetic profiling based on 10 established genetic variants yielded a lower discriminative accuracy for the prediction of CHD (AUC 0.59) than the PROCAM and Framingham risk prediction scores that are commonly used in general populations to guide medical decisions about who should receive risk-reducing medication.[19,20] The AUCs of these conventional risk prediction scores were originally reported to be between 0.75 and 0.82[20,21] but lower in subsequent studies (AUCs ranging between 0.61 and 0.68).[22,23] To obtain AUCs similar to those of the original studies, a considerable number of additional common genetic variants with high ORs needs to be identified. This may prove difficult because the genetic variants with the highest ORs may already have been discovered, which means that even a higher number of genetic variants with small ORs is needed. We are aware that the more extreme scenarios presented in this study might not be realistic and consider our simulated series of 100 genetic variants with ORs of 1.5 and a frequency of 50% each as an upper bound of the plausible range of genetic effects.

The discriminative accuracy that is required in preventive or clinical care depends on the goal of testing, the burden of disease, the availability of (preventive) treatment, and the adverse effects of false-positive and false-

negative test results. For instance, the identification of individuals at increased CHD risk will result in prescription of medication that partly reduces this CHD risk and has a low risk of adverse effects. Here, a small proportion of false positives and/or false negatives is acceptable. Therefore, an AUC between 0.70 and 0.80 would, in this example, indicate moderate performance, whereas an AUC above 0.80 would denote good performance. In contrast, the AUCs needed for decisions about invasive and irreversible interventions or for presymptomatic diagnosis are much higher.

A valuable predictive test should not only have an appreciable AUC but should also be well calibrated. Calibration reflects the level to which the observed probabilities equal the predicted probabilities but is not indicative of the degree to which a test discriminates between individuals who will and will not develop CHD. In other words, a well-calibrated prediction model could show low discrimination. Nevertheless, the calibration was 100% in our simulations because of the way we assigned disease status to the individuals. Recently, the value of the AUC as a measure of the discriminative accuracy has been questioned.[24] Critics claim that researchers should not solely rely on the change in AUC to evaluate the value of risk factors for the prediction of complex diseases but propose to determine the extent of reclassification in clinically important risk categories.[24] Despite the fact that individuals might be reclassified into different risk categories with consequent changes in treatment decisions, the question remains if this reclassification is accurate and correct. We believe, however, that reclassification could be useful in choosing the "best" prediction model when the AUCs of these models are comparable.

An advantage of genetic profiling in predicting CHD risk could be that its information can be used at a younger age before the conventional risk factors, such as hyperlipidemia, hypertension, and diabetes mellitus, become apparent. This could potentially increase the opportunities for prevention of CHD because damaging effects of long-term exposure to conventional risk factors can be prevented. For instance, young individuals who have an adverse genetic profile may adopt a healthier lifestyle. It is, however, not known how likely lifestyle changes will be made and maintained throughout life. An alternative scenario is that genetic profiling may improve the prediction of CHD over and above that of conventional risk factors. In the present simulations, we did not take these conventional risk factors into account, but because of the modeling strategy, our results can be interpreted as combinations of nongenetic and genetic risk factors. Several empirical studies in CHD and other complex diseases have demonstrated limited added predictive value so far.[9,10,25,26] This could be due to the fact that most genetic association studies have focused on genetic variants related to conventional risk factors

for CHD.[5] It is expected that such genetic variants do not have additional predictive value, if the conventional risk factors are already present in the prediction model.[27] In a previous study on CHD risk, a cardiovascular risk score based on conventional risk factors yielded an AUC of 0.76 in whites.[10] The AUC improved to 0.77 by adding 11 genetic risk factors to the prediction model.[10] Such a small improvement, whereas statistically significant, is most likely not clinically relevant.

Our simulation study showed that genetic profiling of 50 or 100 genetic variants could result in higher AUCs compared with the conventional prediction models. Yet, calculation of the proportion of explained variance, as a proxy of the heritability, indicates that some of these scenarios might not be realistic. Twin studies have shown that the heritability of CHD death is in the order of 40%.[28] With a proportion of explained variance of 33%, the AUC was 0.90. Higher proportions of explained variance, yielding higher AUCs, might not be possible for CHD. AUCs higher than 0.90 can only be expected if all genetic variants underlying CHD risk would be identified, including very infrequent genetic variants with very small effects. This may be practically impossible, even with consortia of considerable size. For example, detecting ORs of 1.05, 1.02, or 1.01, whereas 1% of the population has the genetic variant, and with 80% power, requires sample sizes more than 1.3, 8.1, or 32 million individuals.

In conclusion, we have shown that the degree to which genetic profiling is able to predict CHD risk is limited but can be improved by using a large number of common genetic variants with relatively high ORs. To date, we are not able to achieve discriminative accuracies similar to the current CHD risk prediction tools based on the currently known genetic susceptibility variants for CHD. With the latest developments in genomewide association studies, it is expected that the knowledge about genetic risk factors for CHD will expand.[2,29] However, the fact that only one locus identified in the latest genomewide association studies has been successfully replicated is not very promising.[30] The question therefore remains whether the number of additional genetic variants, together with their ORs and genotype frequencies, will be high enough to substantially influence the discriminative accuracy for CHD risk prediction.

References

1. Watkins H, Farrall M. Genetic susceptibility to coronary artery disease: from promise to progress. Nat Rev Genet 2006;7:163-73.
2. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443-53.
3. McCarthy JJ, Parker A, Salem R, et al. Large scale association analysis for identification of genes underlying premature coronary heart disease: cumulative perspective from analysis of 111 candidate genes. J Med Genet 2004;41:334-41.
4. Damani SB, Topol EJ. Future use of genomics in coronary artery disease. J Am Coll Cardiol 2007;50:1933-40.
5. Drenos F, Whittaker JC, Humphries SE. The use of meta-analysis risk estimates for candidate genes in combination to predict coronary heart disease risk. Ann Hum Genet 2007;71:611-9.
6. Yang Q, Khoury MJ, Friedman J, et al. How many genes underlie the occurrence of common complex diseases in the population? Int J Epidemiol 2005;34:1129-37.
7. Humphries SE, Yiannakouris N, Talmud PJ. Cardiovascular disease risk prediction using genetic information (gene scores): is it really informative? Curr Opin Lipidol 2008;19:128-32.
8. Janssens AC, Pardo MC, Steyerberg EW, et al. Revisiting the clinical validity of multiplex genetic testing in complex diseases. Am J Hum Genet 2004;74:585-8.
9. Humphries SE, Cooper JA, Talmud PJ, et al. Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. Clin Chem 2007;53:8-16.
10. Morrison AC, Bare LA, Chambless LE, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol 2007;166:28-35.
11. Janssens AC, Aulchenko YS, Elefante S, et al. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 2006;8:395-400.
12. Peto J. Breast cancer susceptibility—a new look at an old model. Cancer Cell 2002;1:411-2.
13. Pharoah PD, Antoniou A, Bobrow M, et al. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 2002;31:33-6.
14. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown and Company; 1985.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
16. Harrell Jr FE, Shih YC. Using full probability models to compute probabilities of actual interest to decision makers. Int J Technol Assess Health Care 2001;17:17-26.
17. Ash A, Shwartz M. R2: a useful measure of model performance when predicting a dichotomous outcome. Stat Med 1999;18:375-84.
18. Ihaka R, Gentleman RR. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299-314.
19. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). JAMA 2001;285:2486-97.
20. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation 2002;105:310-5.
21. Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.
22. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. Atherosclerosis 2005;181:93-100.
23. Empana JP, Ducimetiere P, Arveiler D, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. Eur Heart J 2003;24:1903-11.

24. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928-35.

25. Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 2008; 358:1240-9.

26. Weedon MN, McCarthy MI, Hitman G, et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 2006;3:e374.

27. Janssens AC, van Duijn CM. Genome-based prediction of common diseases. Hum Mol Genet 2008;17:R166-73.

28. Zdravkovic S, Wienke A, Pedersen NL, et al. Genetic influences on angina pectoris and its impact on coronary heart disease. Eur J Hum Genet 2007;15:872-7.

29. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science 2007; 316:1488-91.

30. Schunkert H, Gotz A, Braund P, et al. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary heart disease. Circulation 2008;117: 1675-84.