Coronary Heart Disease Risk Prediction in the Era of Genome-Wide Association Studies
Current Status and What the Future Holds

Steve E. Humphries, PhD, FRCP, FRCPath; Fotios Drenos, PhD; Gie Ken-Dror, PhD; Philippa J. Talmud, PhD, DSc, FRCPath

For DNA-based tests that assess genetic predisposition to coronary heart disease (CHD) to be of clinical value, they need to provide information over and above conventional risk factors (CRFs) currently used in risk algorithms, such as the Framingham Risk Score,1 which incorporates CRFs such as age, gender, blood lipid concentrations, blood pressure, body mass index, family history, and smoking habit. To achieve this, several hurdles must be passed.

The first challenge is to identify a set of common single-nucleotide polymorphisms (SNPs) at loci associated with CHD risk. Over the last 10 to 15 years, this has been done by use of a “candidate gene” approach through association studies in prospective analysis or case-control studies, ie, comparing SNP genotype or allele frequency between groups of individuals with CHD and healthy subjects. Several of the genes, chosen because of their key role in processes that predispose to atherosclerosis, have meta-analysis–confirmed effects on risk of CHD,2 the best example of which is the APOE gene, which encodes apolipoprotein E, with 3 common isoforms that are associated with strong effects on plasma lipids and more modest effects on risk of CHD.3 This “hypothesis-driven” search for useful genetic variants provides the foundation for the development of genetic CHD risk profiles, and in the last 2 years, it has been enhanced by technical advances that have allowed “hypothesis-free” genome-wide association studies (GWASs), primarily in a case-control setting. Although the list of identified CHD-risk loci and SNPs will clearly grow, we have at least the basis to start the examination of their potential clinical utility.

The second set of challenges is to obtain a robust estimate of the size of the risk effects associated with these SNPs. This requires population-based prospective studies to avoid bias, because estimates in the case-control setting, although efficient for gene discovery, are a suboptimal design for the evaluation of predictive performance of a marker. In addition, information is needed on the risk-allele prevalence between countries and by race and ethnicity, as well as any differences in risk-effect size in these different groups and whether the effect is modified by gender or by the presence of other genes or environmental factors (ie, the context dependence of the effect). To achieve this, genotyping of large data sets will be required, and robust estimates will require that data be combined from several different studies.

Third, the clinical utility of adding these genetic risk scores to the CRF algorithms must be examined, in most cases by use of a simplistic additive model, and the most appropriate clinical setting for its application must be explored given concerns about the psychological impact of DNA testing and confidentiality issues. The final set of challenges, given that the SNPs in the currently identified loci do not represent the full heritability estimate for CHD risk, involves determining how newly emerging data from post-GWAS research can be incorporated into risk algorithms. These research areas include the ability to identify rare or private mutations and the possible utility of measuring telomere length, specific copy number variations (CNVs), or epigenetic effects that result from DNA methylation.

Already, some commercial companies are offering CHD risk genotyping over the Internet, based on the success of the GWASs; for the public to have access to this information is, we believe, premature, and in isolation from CRFs, it has little value. There are still few solid data with regard to stages 2 and 3, and the potential clinical utility of post-GWAS research, outlined in stage 4, is still very unclear. In this review, we examine the progress made in tackling these 4 challenges and discuss the research pathways that may make genetic testing for CHD potentially more likely; the issue of confidentiality and motivation associated with DNA testing is addressed briefly.

Candidate-Gene SNPs Singly and in Combination for CHD Risk Prediction: Modeling Studies

The candidate-gene approach to validate potential CHD risk genes is based on genes chosen because of their involvement...
in CHD-related (patho)physiology and metabolism, and although this approach has been questioned since the development of GWASs, in studies that are adequately powered, it still has validity. However, the use of single SNPs, underpowered studies, population stratification, and lack of replication has led to inconsistencies and poor reproducibility of results, which rightly resulted in candidate-gene studies getting “bad press.” As expected with any single genotype in a multifactorial disease like CHD, published risk estimates have been modest, in the range of 1.12 to 1.75 (reviewed in Casas et al”). Meta-analyses provide the possibility of more robust odds ratios (ORs), as illustrated by the meta-analyses of 15 SNPs in 12 genes that incorporated data from as many as 53 studies. The summary ORs of these meta-analyses were similar, ranging from 0.8 to 1.34. Meta-analyses are also not without bias, including publication bias, population stratification, lack of genotype blinding, control selection bias, and genotype errors, so that even with significant summary ORs, results should be viewed with caution. However, given statistically robust proof of CHD risk, the question arises of how best to use the genotype information in risk algorithms. One approach is to utilize the ORs from meta-analyses to derive a genetic risk function using a binomial distribution for the risk coefficient of each gene, weighted for the frequency of each risk allele. To illustrate this approach, we recently constructed an in silico risk model using 11 SNPs in 10 candidate genes (APOB, eNOS, APOE, ACE, PAI1, MTHFR, GPIIb-IIIa, PON1, LPL, and CETP) and the published (white) allele frequencies and predicted summary risk estimates from meta-analyses. Compared with those with 3 or 4 risk-associated alleles (50% of individuals), those with 6-allele (8.3%) and ≥7-allele (3%) risk genotypes had a significantly higher mean OR for CHD risk (OR 1.70, 95% CI 1.40 to 2.08). This was associated with a steep gradient of CHD risk from those with 6 risk alleles (2.8% of total) to a CHD risk ratio of 3.05 (95% CI 2.32 to 4.14). When age and the risk alleles carried were taken into account, the mean 10-year probability of developing CHD for a 55-year-old man in the lowest decile of genetic risk-associated alleles (50% of total) was 1.55 (95% CI 1.02 to 2.37). Using this method and 12 SNPs in a similar group of genes as reported above, Yiannakouris et al observed that in a subset of the Greek component of the European Prospective Investigation Into Cancer and Nutrition (EPIC) case-control study, the mean gene score was higher in case subjects than in control subjects (P<0.002), and the OR for myocardial infarction associated with a score ≥3.0 (54.3% of the control subjects) was 1.55 (95% CI 1.02 to 2.37).

Candidate-Gene SNPs in Combination: The Risk-Score Approach

The alternative approach to combining SNP information is simply to assign a risk value of, for example, 0 if a subject is a noncarrier of a risk allele, 0.5 or 1 if a carrier, and 1 or 2 if homozygous for that allele, and then to calculate the overall score for each individual. Using this method and 12 SNPs in a similar group of genes as reported above, Kathiresan et al determined that risk prediction improvement will be achievable if enough genetic variants can be identified.

**Figure 1.** Graph of observed LDL/HDL ratio and CHD event rate in subjects from the Malmö Prospective CHD Study. For LDL-C, trend P<10. For HDL cholesterol, trend P<10. Adapted from Kathiresan et al. HR indicates hazard ratio; pyyears, person-years.

Using this gene-score approach, Kathiresan et al determined 9 candidate SNP genotypes, 6 in genes involved in determining levels of low-density lipoprotein (LDL) cholesterol (LDL-C; APOB, APOE, HMGCR, LDLR, PCSK9, and ABCA1) and 3 involved in high-density lipoprotein (HDL) cholesterol (CETP, LIPC, and LPL), with an individual’s score ranging from 0 to 18 for LDL raising or HDL lowering alleles. These were examined in 2246 men with 137 CHD events and 3168 women with 101 events who participated in the Malmö Prospective CHD study and who were followed up for 10.6 years. As shown in Figure 1, the mean LDL/HDL ratio varied from a CHD-protective value of 2.5 in those with >6 risk alleles (2.8% of total) to a CHD risk ratio value of 3.4 in those with >13 risk alleles (5.4% of total). This was associated with a steep gradient of CHD risk from 2.7 to 11 CHD events per 1000 patient-years. Kathiresan et al then examined the ability of the SNPs to discriminate case subjects from control subjects. The AROC value of the CRFs was not improved significantly by the addition of the gene score, but risk classification did improve. Of those at
the reported size of the risk effect are shown in Table 1. Interestingly, although the CELSR2/SORT1/PSRC1 locus is associated with LDL-C levels, 30,31 of which variation in SORT1 appears to be the most likely candidate, 32 other loci show no association with measured phenotypes, and it appears that their mechanism of risk does not operate by influencing known CRF traits such as lipids or blood pressure; thus, the addition of SNPs such as these to the Framingham risk-score algorithm has the potential to improve its overall utility.

The risk-effect sizes found in these GWASs are of the same magnitude (1.2 to 1.6) as those seen in the meta-analyses of candidate SNPs. 33 As for the candidate-gene SNPs, replication of effects and meta-analyses of published GWAS SNP data are vital to validate the findings, even though the original GWAS reports included replication studies (often several of them). To detect these modest effects with a reasonable degree of statistical certainty, very large replication studies are required; for example, even in a combined cohort of 33,382 subjects with 14,366 CHD events, statistically significant risk effects were only confirmed for the GWAS SNPs on chromosomes 1p, 1q, 9p, 10q, and 19q, whereas those on chromosomes 2p, 6q, and 15q had more modest effects that were not statistically significant. 33 Despite the large number of individuals genotyped, these differences may reflect a genuine degree of heterogeneity and context dependency, but it may also be that the original studies identified loci that they were only marginally powered to detect, and these findings would be unlikely to be replicated in other similarly powered studies. Finally, it is also possible that the original findings were spurious owing to a type I error. Achievement of the numbers required to confirm or refute these modest effects will be possible with the establishment of international consortia. Over the next year, there will be genotype data from >200,000 individuals genotyped with the HumanCVD BeadChip (Illumina, Inc, San Diego, Calif), which includes many of the CHD GWAS hits, and this will also provide a source of replication. 34

The GWAS-identified loci reported in Table 1 only represent the locus nearest to the risk-associated SNP(s) and are not necessarily the actual genes involved, and to date, the actual risk-causing variants for any of these loci have not been identified. However, if the linkage disequilibrium or level of correlation between the risk SNP and the disease-causing DNA change is high (eg, >90%), then that SNP will be a good surrogate marker for the variant and could be used, even at this early stage, in a genetic risk-score algorithm. Clearly, further work is needed to identify the functional variants, because their inclusion in the risk algorithm, instead of simply a marker in linkage disequilibrium, will improve risk prediction and reduce uncertainty. One of the current challenges in molecular genetics is to identify these functional variants.

The chr9p21 risk SNPs lie in a gene-poor region, and the nearest genes (CDKN2A-ARF-CDKN2B) are >100-kilobases (kb) upstream from the risk SNPs (Figure 2C). 13-15 CDKN2A/CDKN2B are involved in cell cycle control, and thus, alterations in their expression could be postulated to lead to senescence and apoptosis, both of which are processes...
involved in plaque progression and rupture. However, more detailed analysis of the region between these genes and the risk SNPs has suggested an alternative candidate. In the PROCARDIS study (Precocious Coronary Artery Disease), susceptibility to coronary artery disease was encoded by 2 common haplotypes that span the 53-kb region that overlaps with \textit{ANRIL}, a gene that encodes an antisense noncoding RNA.\textsuperscript{35} This is a member of a gene family involved in transcriptional control that overlaps and regulates \textit{CDKN2B}\textsuperscript{36} and is expressed in atheromatous human vessels in vascular endothelial cells, monocyte-derived macrophages, and coronary smooth muscle cells, all of which are involved in atherosclerosis.\textsuperscript{35} A recent paper (using a mouse model with 58 kb of chr 9p21 deleted) has presented data which suggested that ANRIL expression is not the most likely mechanism, and identified a cis-acting element that influenced expression of \textit{CDKN2A/2B} and thus cell apoptosis.\textsuperscript{35a} This chr9p21 region is clearly a disease “hot spot,” being associated with risk of heart failure,\textsuperscript{37} type 2 diabetes mellitus,\textsuperscript{38} abdominal aortic aneurysms,\textsuperscript{39} stroke,\textsuperscript{37} and periodontal disease,\textsuperscript{40} whereas deletion of this whole region is implicated in certain cancers.\textsuperscript{41} The fact that common variation in this gene region is involved in such a wide range of diseases suggests that this locus encodes 1 or more key players in cell homeostatic processes that are involved in this set of complex multifactorial diseases and raises the possibility that influencing the expression at this locus may have important therapeutic consequences.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Meta-analysis per allele ORs for CHD risk for (A) rs10757274 and (B) rs1333049. The linkage disequilibrium between the SNPs was estimated from data in Samani et al\textsuperscript{26} and kindly made available by the authors as $D' = 0.96$ and $r^2 = 0.88$ in unrelated subjects (Table 5 in Meng et al\textsuperscript{27}). C, Cartoon of the chr9p21.3 locus showing the nearest genes (MTAP, ANRIL, CDKN2A and CDKN2B) and the location of the 2 GWAS-identified SNPs, rs10757274 and rs1333049, reported in the meta-analyses, with a Haploview linkage disequilibrium plot of the region around the 2 SNPs. ARIC indicates Atherosclerosis Risk In Communities; OHS1, OHS2, and OHS3, Ottawa Heart Study 1, 2, and 3; CCHS, Copenhagen City Heart Study; DHS, Dallas Heart Study; FH, familial hypercholesterolemia; WGHS, Women’s Genome Health Study; WTCCC, Wellcome Trust Case Control Consortium; MI, myocardial infarction; GerMIFS II, German Myocardial Infarction Family Study II; UK MI, United Kingdom Myocardial Infarction Study; MONICA/KORA, Monitoring of Trends and Determinants in Cardiovascular Disease/Cooperative Health Research in the Region Augsburg; PRIME, Prospective Epidemiological Study of Myocardial Infarction; and AMC-PAS, Academic Medical Center Amsterdam Premature Atherosclerosis Study.}
\end{figure}
In Silico Modeling of Effect of Combined Risk SNPs

We have modeled the use of 7 of these “novel” GWAS SNPs in the risk-score algorithm in combination with the 11 candidate-gene SNPs discussed above using the published risk-allele frequency in whites. The predicted distribution of individuals with different numbers of the combined 17 risk alleles is shown in Figure 3. The most common group will have 5 risk alleles, with 10% having only 3 and 5% having 2 or fewer, whereas 6.8% have 8 risk alleles, 2.7% have 9, and 1% have 10 or more. When the reported estimates of CHD risk per allele are used, compared with risk in the commonest group, those with 3 risk alleles have a roughly 33% lower and those with 2 or fewer risk alleles have 50% lower CHD risk. By contrast, those with 8 risk alleles have a risk of >1.9, those with 9 have a risk approaching 2.5, and those with 10 have a risk of >3.1; by simulation in 10,000 subjects, these effects of 2 or fewer and 10 or more risk alleles are statistically significant (online-only Data Supplement). It is of relevance that the risk associated with being a current cigarette smoker is roughly 2-fold that of a nonsmoker, and so we can predict from this that 2% to 3% of the population will be at a clinically important higher genetic risk and roughly 4% will be at lower risk on the basis of the combined information from these SNPs. Although it is clear that the risk of smoking can be reduced by quitting, there are appropriate interventions for those who have a high genetic risk profile, such as more aggressive treatment of modifiable risk factors such as plasma cholesterol levels and blood pressure by pharmacological intervention, lifestyle interventions to re-

Table 1. List of Other Chromosomal Localizations Associated With CHD or MI From GWAS That Have Been Replicated in More Than 1 GWAS

<table>
<thead>
<tr>
<th>Chromosome Localization/Gene</th>
<th>SNP and Risk Genotype</th>
<th>Risk Allele Frequency</th>
<th>Size of the Effect (95% CI)</th>
<th>Overall P</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13.3 CELSR2/PCSR1/SORT1</td>
<td>rs646776-T*</td>
<td>0.81</td>
<td>1.19 (1.13–1.26)</td>
<td>7.9 × 10^{-12}</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>rs599839-A†</td>
<td>0.28</td>
<td>1.13 (1.08–1.19)</td>
<td>1.4 × 10^{-7}</td>
<td>26</td>
</tr>
<tr>
<td>1q41 MIA3</td>
<td>rs17465637-C*</td>
<td>0.72</td>
<td>1.14 (1.10–1.19)</td>
<td>1.4 × 10^{-9}</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>rs3008621-G†</td>
<td>0.16</td>
<td>1.10 (1.04–1.17)</td>
<td>1.0 × 10^{-3}</td>
<td>26</td>
</tr>
<tr>
<td>2q36</td>
<td>rs2943634-C†</td>
<td>0.34</td>
<td>1.05 (1.0–1.1)</td>
<td>0.03</td>
<td>26</td>
</tr>
<tr>
<td>3q22.3 MRAS</td>
<td>rs9818870-T†</td>
<td>0.20</td>
<td>1.15 (1.11–1.19)</td>
<td>7.4 × 10^{-13}</td>
<td>29</td>
</tr>
<tr>
<td>6q25 MTHFD1L</td>
<td>rs6922269-A*</td>
<td>0.26</td>
<td>1.09 (1.0–1.1)</td>
<td>2.3 × 10^{-5}</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>rs6922269-A†</td>
<td>0.26</td>
<td>1.05 (1.0–1.1)</td>
<td>0.02</td>
<td>26</td>
</tr>
<tr>
<td>10q11 CXCL12</td>
<td>rs1746048-C*</td>
<td>0.84</td>
<td>1.17 (1.11–1.24)</td>
<td>7.4 × 10^{-9}</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>rs501120-T†</td>
<td>0.13</td>
<td>1.11 (1.05–1.18)</td>
<td>4.3 × 10^{-4}</td>
<td>26</td>
</tr>
<tr>
<td>12q24 HNF1A/C12orf43</td>
<td>rs2259816-A‡</td>
<td>0.36</td>
<td>1.05 (1.05–1.11)</td>
<td>4.8 × 10^{-7}</td>
<td>29</td>
</tr>
<tr>
<td>15q22 SNAAD3</td>
<td>rs17228212-T*</td>
<td>0.73</td>
<td>1.05 (1.01–1.09)</td>
<td>0.02</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>rs17228212-C†</td>
<td>0.26</td>
<td>1.00 (0.95–1.04)</td>
<td>0.9</td>
<td>26</td>
</tr>
</tbody>
</table>

MI indicates myocardial infarction.

Data include original discovery study and replication studies.
*From overall analysis in 12,713 case subjects and 12,821 control subjects.28
†From overall analysis in 11,550 case subjects and 11,205 control subjects.26
‡From overall analysis in 19,407 case subjects and 21,366 control subjects.29

In Silico Modeling of Effect of Combined Risk SNPs

We have modeled the use of 7 of these “novel” GWAS SNPs in the risk-score algorithm in combination with the 11 candidate-gene SNPs discussed above using the published risk-allele frequency in whites. The predicted distribution of individuals with different numbers of the combined 17 risk alleles is shown in Figure 3. The most common group will have 5 risk alleles, with 10% having only 3 and 5% having 2 or fewer, whereas 6.8% have 8 risk alleles, 2.7% have 9, and 1% have 10 or more. When the reported estimates of CHD risk per allele are used, compared with risk in the commonest group, those with 3 risk alleles have a roughly 33% lower and those with 2 or fewer risk alleles have 50% lower CHD risk. By contrast, those with 8 risk alleles have a risk of >1.9, those with 9 have a risk approaching 2.5, and those with 10 have a risk of >3.1; by simulation in 10,000 subjects, these effects of 2 or fewer and 10 or more risk alleles are statistically significant (online-only Data Supplement). It is of relevance that the risk associated with being a current cigarette smoker is roughly 2-fold that of a nonsmoker, and so we can predict from this that 2% to 3% of the population will be at a clinically important higher genetic risk and roughly 4% will be at lower risk on the basis of the combined information from these SNPs. Although it is clear that the risk of smoking can be reduced by quitting, there are appropriate interventions for those who have a high genetic risk profile, such as more aggressive treatment of modifiable risk factors such as plasma cholesterol levels and blood pressure by pharmacological intervention, lifestyle interventions to re-
duce obesity and stress, and modification of diet and alcohol intake. Although not yet proven, it is reasonable to assume that these risk factor modifications will be equally effective in subjects at genetically high risk as in subjects in the general population, as has been found, for example, with the reduction in CHD mortality seen with lipid-lowering therapies in patients with the monogenic disorder of familial hypercholesterolemia. However, although it appears very likely that these SNPs will add to risk prediction over and above the CRF-based Framingham risk score, further data are required to confirm this.

**Clinical Utility of Genotype Risk Stratification: Chromosome 9p21 SNP as an Example**

Prospective studies of incident as opposed to prevalent CHD are required to assess the utility of both biomarkers and genotypes with regard to CHD prediction. Case-control studies are efficient for gene discovery, but because they usually recruit more severely diseased or younger patients, they are unrepresentative of the disease cases in the general population, they may provide biased information on population allele frequency, and they cannot give unbiased estimates of attributable risk or the effect of genes on other important risk factors for cardiovascular disease (CVD). Most importantly, only prospective studies allow an accurate estimation of the absolute risk associated with a genotype or permit evaluation of the predictive utility of genetic information over and above the impact of classic risk factors. We therefore investigated whether the addition of a chr9p21 SNP genotype improved the prediction of CHD events by CRFs in the Framingham risk-score algorithm in 2742 healthy middle-aged men from the prospective Northwick Park Heart Study II (NPHSII) who were followed up for an average of 14 years with 270 CHD events.\(^\text{17}\) The rs10757274 G allele was associated with incident CHD with similar effect size to that observed in case-control studies, (hazard ratio of 1.60 in GG compared with AA men, 95% CI 1.12 to 2.28). The population-attributable fraction for CHD explained by the SNP was 26.2% (95% CI 7.1 to 41.1), which was independent of CRFs and reported family history of early CHD. Although this high population-attributable fraction estimate would suggest that this variant is likely to be of clinical utility, this is not an appropriate statistic measure to demonstrate this. To examine this, the ability to risk stratify, or “discrimination,” was evaluated by the \(A_{\text{ROC}}\), and when genotype was added to the model, perhaps surprisingly given its large effect in univariate analysis, the \(A_{\text{ROC}}\) for CRFs alone of 0.62 (95% CI 0.58 to 0.66) was nonsignificantly (\(P=0.14\)) increased by 3% to 0.64 (95% CI 0.60 to 0.68). Similarly, in the Women’s Genome Health Study of 22 129 white women followed up for \(\approx10\) years with 715 CVD events,\(^\text{18}\) the same chr9p21 SNP was significantly associated with CHD risk in univariate analysis (hazard ratio in carriers 1.25, 95% CI 1.04 to 1.51), but the addition of this genotype to a CRF-based algorithm did not improve the \(A_{\text{ROC}}\). However, it is now increasingly recognized that prediction can only be improved significantly by the inclusion of factors that are both common and have very large effects,\(^\text{19}\) and that a single genotype (or biomarker) associated with ORs in the region of 1.2 to 1.6 will not on its own significantly improve risk prediction for polygenic, multifactorial CHD. To improve an already high \(A_{\text{ROC}}\), biomarkers (or single SNPs or SNPs in combination) with large effects that are also common are needed, and there is diminishing added value in the higher \(A_{\text{ROC}}\) ranges for variants with similar effects, as we demonstrated in modeling the added effects of SNPs with similar frequency and risk size as the chr9p21 variant.\(^\text{17}\) We modeled the effect on CHD risk of up to 10 hypothetical, randomly assigned gene variants with allele frequencies and risk similar to those of rs10757274. The addition of 1 further SNP with similar characteristics increased the \(A_{\text{ROC}}\) significantly (\(P<0.03\)), whereas the inclusion of 2 or more SNPs had a greater effect (\(P<0.001\)), with the addition of further SNPs having smaller incremental effect. The \(A_{\text{ROC}}\) for 10 SNPs was 0.76.\(^\text{17}\)

However, the more relevant clinical use for such genetic information is in risk calibration, which is the ability of the genotype to improve the stratification of individuals into risk categories (ie, whether genotype information increased the number of men correctly reclassified into those developing CHD or remaining event free during follow-up). The potential clinical utility of this genotype for risk stratification in the NPHSII men was also examined. Although treating subjects with a low Framingham risk score will not be cost-effective because of the low event rate, the majority of CHD events occur in subjects with intermediate risk scores, because this is the most common group. The National Institute for Clinical Excellence, which sets standards of clinical treatment in the United Kingdom, recommends that subjects with a 10-year risk of CVD of >20% should be treated with statins (http://www.nice.org.uk/nicemedia/pdf/CG67NICEguideline.pdf).

Using the variables in the Framingham algorithm, this equates to a risk score of 23. As shown in Figure 4, of the 2670 men with a chr9p21 genotype and complete trait data, 164 were over the score, of whom 33 had an event (observed 10-year risk 23.3%). With the addition of the rs10757274 SNP, there were 55 men who were originally below this cutoff on the basis of CRFs alone (2.1% of the whole group who would not have qualified for statin treatment) but who would have a score >23 if rs10757274 genotype were added, and of these, 12 men went on to experience a definite CHD event (observed 10-year risk 24.0%). After the inclusion of the genotype, 585 men (21.9%) were reclassified, of whom 63% (369) moved into more accurate categories (defined as those above the 20% 10-year disease-risk cutoff were 6.73 and 6.78 mmol/L, respectively (baseline LDL-C levels 4.2 and 4.8 mmol/L, respectively), and based on the expected benefit of reducing their individual
Post-GWAS DNA-Based CHD Risk Factors

Although the results from the GWASs for CHD and CHD risk factors were originally viewed as very exciting, there has been some disappointment expressed recently that the identified genes and their common variants (as detected by the tagging SNPs) explain only a small proportion of the population risk of CHD or the levels of its risk-associated phenotypes (for example, LDL-C levels). The commercial GWAS SNP platforms do not give complete coverage of the genome, with certain regions poorly covered (for example, the APOE and LPL loci), so that the influence of variation at these regions in determining lipid levels and CHD risks is very likely to have been underestimated, and some novel loci may have been missed completely. The combining of data from many GWASs will enable genes of smaller effect to be identified with a greater degree of statistical certainty, and although it appears unlikely that any SNPs with large effects on CHD risk (relative risk of $>1.2$) remain to be found, this is likely to lead to the identification of many additional CHD risk loci with ORs in the range of 1.1 to 1.15, although larger effects in subgroups, eg, smokers, people with diabetes mellitus, or different ethnic groups, may be revealed. Therefore, it seems probable that the current set of genes (with refinements with respect to the best SNPs for each gene) is likely to form the foundation of any DNA-based genetic testing algorithm, with the list being supplemented by additional SNPs of smaller effect. Further predictive information may be achievable if any examples of gene-gene interaction can be identified robustly, ie, the effect on CHD risk of an individual carrying 2 particular risk alleles of different genes is significantly greater than the sum of their individual effects. To date, no such examples have been identified robustly, but again, the combining of data from several consortia in the next few years will lead to the ability to confirm or refute any examples. Currently, models use additive effects only of SNPs and CRFs, but with larger data sets, it should be possible to explore the possibility that more complex models of combined effects fit the data better and thus give better prediction.45

Future Prospects: Novel Genetic and Epigenetic CHD Risk Information

There are several potentially relevant fields of research in which additional DNA-based information has the potential to add to an individual’s CHD risk profile. These findings all require further years of research before they reach the level of certainty required to be included in clinical risk assessment.

Very Rare or Private Mutations

Although the common variants (ie, a minor allele frequency of $>1\%$) identified in the published GWASs for CHD risk...
and CHD traits only explain a small percentage of the disease and trait variance; they clearly identify at least some of the relevant genes. Thus, a complementary approach to explain an individual’s CHD risk would be to use sequencing to identify rare variants in these same genes. This approach is exemplified by the gene that encodes the LDL receptor (LDLR), in which common variants have a modest effect on LDL-C levels and rare variants in the gene, found in subjects with familial hypercholesterolemia, cause very large elevations of LDL-C and a marked CHD risk. With the development of third-generation sequencing technologies, it is now possible to obtain a huge amount of genomic sequence information from an individual relatively inexpensively, and this will become faster and cheaper in the very near future.

The most important advance is the ability to use “capture” techniques to focus on a range of specific genes and targeted regions, such as promoters, exons, and intron-exon junctions, so that key genes for CHD can be examined in an efficient manner. This uses either custom-designed or commercially available “whole-exon” high-density oligonucleotide microarrays to “capture” by hybridization short genome segments, for example, that include all of the individual exons of a gene or series of genes, or even complete long regions that correspond to entire gene loci, which are then sequenced.

Several studies that have used this approach for genes involved in CHD traits, such as HDL and LDL-C, have reported that a significant proportion of individuals with extreme levels of these lipid traits have sequence changes that are absent from individuals in the general population (or at the other extreme of the trait). On the face of it, this information would appear to be valuable for the general population (or at the other extreme of the trait). On the face of it, this information would appear to be valuable for identifying rare variants in these same genes. This approach is relevant genes. Thus, a complementary approach to explain the likelihood of the change affecting function. For changes that potentially affect splicing, although those in the canonical AT/GC regions of intron-exon boundaries can be “called” with a high degree of reliability, changes further into introns are more problematic; again, computer algorithms may be helpful. Finally, the prevalence of the sequence change in ethnically matched healthy subjects is determined, because healthy carriers indicate that the change is not disease causing. In the field of CHD, many laboratories already have good experience in this area because of work in patients with the monogenic disorder familial hypercholesterolemia, but ultimately, definitive proof of causality may require in vitro expression, which is obviously time-consuming and expensive. Additionally, although a familial hypercholesterolemia–causing mutation may be expected to have a relatively large effect on function, a much more modest effect of a gene-variant contribution to polygenic CHD is predicted (eg, a promoter variation that alters gene expression or a coding change that alters enzyme function by 15% to 20%), which will be much harder to detect in vitro with statistical certainty. For example, in vitro expression of 20 “private” NPC1L1 variants showed decreased function in 14, but 6 did not have a significant effect in this assay.

The second issue is what CHD risk to assign the identified variant. As an example, variants have been identified in the ATP-binding cassette transporter-1 receptor gene (ABCA1) that were associated with lower plasma HDL cholesterol levels and in vitro function but not with a significant effect on CHD risk, and similarly, a variant in the hepatic lipase gene, LIPC, that influenced HDL levels was not associated with the expected CHD risk. The identified common functional variants have meta-analysis–estimated effects in the range of 1.2 to 1.4, but robust proof of these effects has required pooling of many studies, and the combined effects of carrying more than 1 such variant still have not been explored. For polygenic CHD, it is currently unclear how a reliable risk estimate could be assigned to the collection of several hundred “private” DNA differences any individual is likely to have compared with the “normal” reference sequence, even if only 8 to 10 genes are examined, for example. Data from such approaches will certainly become available in the future (eg, from the “1000 Genomes” project, http://www.1000genomes.org) and may help overcome some of these problems.

Telomere Length
Telomeres are specialized DNA-protein structures at the end of all chromosomes that preserve chromosome stability and integrity. In humans, they are made up of variable numbers of tandem repeats of a TTTAGGG sequence. Lengths show a wide range of interindividual variability, but twin studies have indicated a heritability of ~80%. Telomeric DNA is incompletely replicated during cell division, which means that telomeres shorten with every cycle. The action of the telomerase enzyme may partially compensate for this effect in some tissues and/or at different stages of life. Critically, telomere length is important to normal aging, and premature telomere shortening has been implicated in cancer, heart disease, and other degenerative disorders. In support of this, subjects with CHD have mean telomere lengths similar to those of healthy subjects who are approximately 9 years older, which suggests either faster shortening due to risk factors found during the development of CHD (eg, oxidative stress) or shorter lengths at birth in those predisposed to CHD. Compared with subjects in the highest quartile for telomere length, the risk of myocardial infarction for subjects in the lowest quartile was increased approximately 3-fold, an effect that was independent of classic risk factors but was influenced by oxidative stress.
Table 2. Recently Reported “Deep-Sequencing” Results for Candidate Lipid Genes, Samples Used, Prevalence of Identified Variants, and Effect on Traits or CHD Risk

<table>
<thead>
<tr>
<th>Gene Name (Symbol)</th>
<th>Protein Function</th>
<th>Sample Strategy (No. of Samples)</th>
<th>Prevalence of Identified Variants</th>
<th>Effect on Traits/CHD</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angiopetin-like-4 (ANGPTL4)</td>
<td>Plasma inhibitor of lipoprotein lipase–mediated hydrolysis of TG-rich lipoproteins</td>
<td>Low and high quartile of TG (3551)</td>
<td>93 Variants identified, ~50% found in only 1 individual</td>
<td>Prevalence of synonymous and noncoding variants same in low vs high, but frequency of synonymous/truncation mutations higher in high TG group</td>
<td>52</td>
</tr>
<tr>
<td>ATP-binding cassette transporter-1 (ABCA1)</td>
<td>Receptor involved in HDL-mediated cholesterol efflux</td>
<td>Healthy subjects with low (190) and high (190) HDL-C</td>
<td>Tested 7 synonymous variants. Overall carrier frequency of any variant 3/1000</td>
<td>4 Variants were associated with HDL-C and 3 did not. Variants caused 30% to 50% lower cholesterol efflux in vitro. None had a significant effect on CHD risk</td>
<td>48</td>
</tr>
<tr>
<td>Niemann-Pick type C1-like-1 (NPC1L1)</td>
<td>Involved in dietary cholesterol absorption</td>
<td>Subjects with low (256) and high (256) cholesterol absorption</td>
<td>Nonsynonymous variants in 10% of low absorbers vs 2% of high absorbers. Most found in 1 individual only</td>
<td>Plasma LDL-C lower among those with low cholesterol absorption. Of 20 alleles found in low absorbers, 14 had significantly reduced expression in vitro</td>
<td>53, 54</td>
</tr>
<tr>
<td>Hepatic lipase (LIPC)</td>
<td>Enzyme involved in lipoprotein (particularly HDL) remodeling</td>
<td>Previously identified synonymous variants</td>
<td>Tested 4 variants of carrier frequency 5–60/1000</td>
<td>Only 1 (S267F, frequency 7/1000) was associated with 16% higher HDL. None had a significant effect on CHD risk</td>
<td>50</td>
</tr>
<tr>
<td>Endothelial lipase (LIPG)</td>
<td>Enzyme involved in HDL remodeling</td>
<td>Healthy subjects with low and high HDL-C (1170 chromosomes)</td>
<td>10 Nonsynonymous variants (7 unique) in low HDL vs none in high HDL</td>
<td>Computer prediction indicates most are probably damaging. In vitro, every variant caused a decrease in enzymatic activity. T111I had no effect on plasma HDL. N396S caused ~16% higher HDL</td>
<td>49</td>
</tr>
<tr>
<td>Lipoprotein lipase (LPL), apolipoprotein CII (APOC2), and ApoAV (APOA5)</td>
<td>Key enzyme and key apolipoprotein cofactors of plasma triglyceride clearance</td>
<td>High TG patients (110)</td>
<td>6 in LPL (2 novel), 2 in APOC2 (1 novel) and APOA5 (known). Overall prevalence 10%. All absent from control subjects</td>
<td>Computer prediction suggested probable or possible effect on function</td>
<td>51</td>
</tr>
</tbody>
</table>

TG indicates triglycerides; HDL-C, HDL cholesterol.

Although technically challenging, mean telomere length can be estimated by a quantitative polymerase chain reaction method that determines the relative concentration of telomere sequence present compared with a single-copy gene. Because telomere length is most conveniently measured in DNA from circulating leukocytes, the issue of the correlation between telomere length in these cells and in the biologically relevant vascular tissue is important. Wilson et al have shown that vascular wall cell telomere length from patients undergoing elective abdominal aortic repair, and from normal cadavers, strongly correlated with leukocyte telomere length ($r=0.62, P=0.001$). Shorter telomere length thus probably reflects shorter vascular telomere length and suggests a relative “aging” of the vascular wall. If this relationship between shorter telomeres and CHD can be confirmed in prospective studies, it may be a useful additional DNA marker of an individual’s future risk.

**Epigenetic Modification by Methylation**

Epigenetics refers to the transmissible changes in gene expression caused by mechanisms other than a change in the underlying DNA sequence. These changes remain through cell divisions for the entire life of the cell and may also last for multiple generations. The molecular basis of epigenetics is complex but is mainly due to chromatin modification or the addition of a methyl group at the 5′-carbon position of the cytosine residues located in the context of dinucleotide CpG sites. CpG dinucleotides are statistically underrepresented in the genome but are found to be concentrated in C+G–rich regions (CpG islands), usually 500 to 2000 base pairs in length, which are frequently located in and around the transcription start sites of human genes. Methylation of these CpG islands is usually associated with silencing of the respective gene.
Methylation profiles reflect environmental exposure and could thus contribute to CVD risk. An example of this comes from a study of nearly 800 individuals exposed to traffic pollution, which is known to generate particulate matter and which has been associated with CHD risk. Investigation into the methylation status of long interspersed nucleotide element (LINE)-1 showed that there was a relationship between the hypomethylation status of these repeat elements and pollution exposure. Whether this is associated with increased CHD risk has not been determined, but it does suggest a direct relationship between some environmental factors and methylation profile. The importance of epigenetics in CVD has been suggested by the association of aberrant DNA methylation with a predisposition to and a natural history of atherosclerosis. In addition, over the last 20 years, an increasing amount of epidemiological and pathological evidence has become available that illustrates the relationship between an adverse in utero environment and an increased risk of vascular disease in the offspring. DNA methylation appears to mediate the long-term memory and thus developmental programming of cells and tissues, which eventually leads to the disease. With methylation-sensitive restriction enzymes to identify methylation sites, genomic methylation was found to be significantly higher in patients with angiographically defined CHD than in control subjects and showed a strong correlation with homocysteine levels, a key modulator of macromolecular methylation. An individual’s plasma homocysteine levels are determined by both environmental (dietary) and genetic factors, including common functional variants in genes that encode key enzymes in the cysteine metabolism pathway, most notably in methionine tetrahydrofolate reductase (MTHFR). Individuals who are homozygous for a thermolabile amino acid variant in the gene (≈10% of the white population) have higher plasma homocysteine and meta-analysis–proven higher CHD risk.

Again, determining the extent of methylation of particular CpG sites within a gene of interest is technically challenging. There are several commercial chip-based methods available, but this is a rapidly changing field, and within a year, new technologies should be available. Probably the most exciting advance in this area comes from the possibility of combining haplotype and epigenetic information, where it has been shown that allele-specific methylation was recurrently dependent on genotype, or in other words, sequence-dependent, allele-specific methylation.

Copy Number Variations

In the last 5 years, there has been increasing interest in regions of the genome that exhibit interindividual heterogeneity in copy number. These CNVs have been identified by novel genome-scanning technologies, as has their involvement in complex disease phenotypes. CNVs reflect submicroscopic variants that range from 1 kb in size to several megabases and include deletions, duplications, inversions, and complex multisite rearrangements. Differences in copy number may alter gene function, gene dosage, or, perhaps more likely, noncoding sequences that might influence expression levels without affecting gene function. Structural variants in the genome are thought to contribute to approximately 15% of the human genome. CNVs are associated with several common complex disorders (reviewed in Estivill et al). The ubiquity of CNVs is highlighted by the catalogue of almost 20,000 CNVs that correspond to >6,000 loci detailed on the Database for Genomic Variants (http://projects.tcag.ca/variation).

Detection and characterization of CNVs remains complex and somewhat inexact, with the most specific method being array comparative genomic hybridization or analysis of fluorescence intensity from genotyping arrays to give a direct measure of copy number at a given locus. It has also been suggested that linkage disequilibrium between an SNP and a CNV may be exploited, because with sufficiently high linkage disequilibrium, the genotype of the CNV of interest can be predicted with high probability if the genotype of the CNV “tagging” SNP is known. Although not yet robustly associated with CHD or with CHD risk traits, CNVs thus represent a source of genetic variability that could contribute significantly to cardiovascular disease.

DNA-Based Testing, Confidentiality, and Motivation

Full coverage of this topic is beyond the scope of the present review, but a few points should be considered briefly. Although the concept of presymptomatic identification based on biomarkers (currently, classic risk factors) is well established, there is an additional concern about the use of genetic information (more specifically, DNA-based tests) with regard to confidentiality, access to certain insurance products, and discrimination in the workplace. Our belief is that these issues are serious but not insurmountable and that the way forward is to develop appropriate methods of sample and data handling and encrypting, as for any biochemical or other medical information, to give people confidence in the safety of their personal information.

The second issue is whether giving people DNA risk information has a positive or negative influence on their motivation to make the necessary changes to their lifestyle to reduce their subsequent CHD risk. These changes might include weight loss, smoking cessation, and adherence to prescribed medication. It is clear that current approaches to motivating adherence to prescribed medication are poor, with studies indicating that even after a myocardial infarction, adherence to statin use falls from 85% on hospital discharge to 56% at 6 months and 45% at 1 year, with adherence in the case of primary prevention being even lower. If DNA-based risk information improved motivation even by a modest amount, this would be extremely valuable. However, concern has also been raised that DNA information may be viewed fatalistically, and thus, motivation would be reduced. Familial hypercholesterolemia is one situation in which genetic testing is currently feasible, and in a randomized controlled trial of cholesterol only versus cholesterol plus DNA diagnosis, it has been reported that in the relatives of patients with familial hypercholesterolemia, the DNA-based diagnosis was no more worrying than a cholesterol test and had no significant long-lasting effect on psychological well-being. One additional benefit appeared to be that people with a DNA-based diagnosis were keener to use medication to reduce their
cholesterol levels. Whether or not this translates into better adherence is an issue for future research.

However, it is also important to distinguish “disease-causing” genetic testing from “disease-susceptibility” genetic testing. In general, tests for disease-causing mutations are based on single-gene tests for monogenic disorders that have a high penetrance. Tests for Huntington’s disease and for breast cancer are good examples, and in general, these tests are associated with positive psychological changes in both mutation carriers and noncarriers. It is unclear to what extent this finding can be extrapolated directly to tests of disease susceptibility, which are based on the combined information of several or many DNA sequence changes of modest effect. Thus, the impact and potential limitations of both types of tests may be very different. To date, there are very few studies that have examined the psychological or motivational impact of either single or multiple SNP test results for disease susceptibility, although the results suggest that these appear to be generally beneficial; for example, results from a small randomized controlled trial that used an SNP in GSTM1 associated with the risk of lung cancer to motivate smoking cessation supported the view that this information was not perceived fatalistically and may even be motivating. Further research is required using randomized controlled trials to examine the utility of incorporating genetic information into risk assessments for type 2 diabetes mellitus, as well as trials of interventions designed to motivate lifestyle improvements in individuals at risk of developing CHD.

Conclusions

Currently, estimation of an individual’s risk of future CHD risk is most accurately accomplished by conventional measures. However, if the genetic architecture of these key traits can be better understood, this genetic information may have predictive value beyond measures of CRFs when added to the risk algorithm. Currently, studies are needed to determine the impact of the old candidate and new GWAS SNPs on risk in combination and to estimate this impact beyond that of CRFs. Although improving the clinical utility of risk-prediction algorithms is likely to require the availability of a large number of validated SNPs (eg, 40 or more), the data suggest that even at the current state of knowledge, the SNPs we have available will allow clinicians to stratify individuals who are at intermediate risk, for example, and therefore make recommendations about the management of such individuals that are clinically useful. Because many SNPs can be determined inexpensively and simultaneously in a single sample (for example, using DNA obtained from a buccal swab), this will not have major cost implications. Thus, in the future, a patient could be booked for a clinic visit in 2 to 3 weeks and could be sent a mouthwash tube to be returned to the laboratory immediately, so that genetic information would be available for discussion. This sample could easily be tested for 20 to 100 different variations, and with falling costs and high-throughput genotyping, costs in the range of $30 to $80 are achievable, although “deep sequencing,” CNVs, telomere length, and methylation pattern analyses currently require DNA from blood and are much more expensive. Perhaps one of the areas in which research in this field is most urgently needed is the exploration of different ways to present such genetic risk information to individuals so as to find approaches that minimize a sense of fatalism and maximize motivation for behavior change (see Sanderson and Michie). More importantly, an understanding at the molecular level of the mechanisms of the additive and interactive effects between genetic polymorphisms and between genetic and environmental factors is the foundation for developing novel therapeutic strategies.

Acknowledgment

All authors have read and agree to the manuscript as written.

Source of Funding

This work is supported by the British Heart Foundation (PG/2008/0008, SP/07/007/23671, and RG05/014).

Disclosures

None.

References

15. 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;447:661–678.


46. Wilson WR, Herbert KE, Mistry Y, Stevens SE, Patel HR, Hastings RA, Thompson MM, Williams B. Blood leukocyte telomere DNA content...


**Key Words:** genetics ■ risk factors ■ prevention ■ risk assessment ■ genome-wide association studies
Coronary Heart Disease Risk Prediction in the Era of Genome-Wide Association Studies: Current Status and What the Future Holds
Steve E. Humphries, Fotios Drenos, Gie Ken-Dror and Philippa J. Talmud

Circulation. 2010;121:2235-2248
doi: 10.1161/CIRCULATIONAHA.109.914192
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2010 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/121/20/2235

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org//subscriptions/