Human Molecular Genetics, 2009, Vol. 18, No. 18 **3525–3531**
doi:10.1093/hmg/ddp295
Advance Access published on June 24, 2009

Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk

David M. Evans[1,*], Peter M. Visscher[2] and Naomi R. Wray[2]

[1]Department of Social Medicine, MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, UK and [2]Genetic Epidemiology and Queensland Statistical Genetics, Queensland Institute of Medical Research, Australia

Received February 26, 2009; Revised June 15, 2009; Accepted June 22, 2009

The current paradigm within genetic diagnostics is to test individuals only at loci known to affect risk of complex disease—yet the technology exists to genotype an individual at thousands of loci across the genome. We investigated whether information from genome-wide association studies could be harnessed to improve discrimination of complex disease affection status. We employed genome-wide data from the Wellcome Trust Case Control Consortium to test this hypothesis. Each disease cohort together with the same set of controls were split into two samples—a 'Training Set', where thousands of SNPs that might predispose to disease risk were identified and a 'Prediction Set', where the discriminatory ability of these SNPs was assessed. Genome-wide scores consisting of, for example, the total number of risk alleles an individual carries was calculated for each individual in the prediction set. Case–control status was regressed on this score and the area under the receiver operator characteristic curve (AUC) estimated. In most cases, a liberal inclusion of SNPs in the genome-wide score improved AUC compared with a more stringent selection of top SNPs, but did not perform as well as selection based upon established variants. The addition of genome-wide scores to known variant information produced only a limited increase in discriminative accuracy but was most effective for bipolar disorder, coronary heart disease and type II diabetes. We conclude that this small increase in discriminative accuracy is unlikely to be of diagnostic or predictive utility at the present time.

INTRODUCTION

Genetic testing of monogenic diseases where there is a strong correlation between risk genotype and disease has been employed successfully in a diverse range of applications from prenatal and newborn screening, to carrier testing and medical diagnostics (1,2). With the advent of genome-wide association studies (GWAS) and the subsequent identification of well over 150 genetic loci contributing to common complex disease (3), attention has now turned to whether genetic testing could also be applied successfully in diagnosing/predicting complex disease, and in so doing, herald a new era of personalized medicine.

Despite some of the initial enthusiasm regarding the potential of genetic testing in common complex diseases (4–6), much of the excitement has been tempered by the realization that the small effect sizes and the often low to moderate heritabilities of common diseases, mean that the predictive utility of genetic testing is likely to be limited (7,8). A major problem is that the effect sizes of individual alleles are small, typically in the range 1.1–2. The predictive value afforded by a single variant of small effect is therefore likely to be negligible. This has led to the idea of testing multiple genetic loci simultaneously, also called 'genomic profiling', which collectively may result in superior prediction of complex disease (9,10). At the current time, it is too early to say whether genomic profiling will prove to be clinically useful, but preliminary results for some complex diseases such as adult macular degeneration (11) together with simulation studies suggest that genomic profiling may have at least some utility in identifying high risk groups in screening programs (12).

*To whom correspondence should be addressed at: Department of Social Medicine, MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK. Tel: +44 1173310094; Fax: +44 1173310123; Email: dave.evans@bristol.ac.uk

Table 1. Median AUC values for the seven diseases using the count and log odds methods

Threshold	BD	CHD	HT	CD	RA	T1D	T2D
Count method							
0.8	0.653 (0.527)	0.599 (0.527)	0.602 (0.538)	0.617 (0.554)	0.591 (0.522)	0.620 (0.513)	0.589 (0.520)
0.5	0.664 (0.527)	0.598 (0.533)	0.600 (0.534)	0.622 (0.553)	0.594 (0.528)	0.624 (0.515)	0.593 (0.513)
0.1	0.646 (0.537)	0.570 (0.532)	0.587 (0.534)	0.596 (0.524)	0.592 (0.515)	0.637 (0.515)	0.578 (0.516)
0.05	0.625 (0.537)	0.552 (0.525)	0.589 (0.532)	0.591 (0.521)	0.599 (0.524)	0.673 (0.537)	0.576 (0.529)
0.01	0.570 (0.555)	0.588 (0.508)	0.566 (0.518)	0.561 (0.514)	0.625 (0.522)	0.697 (0.531)	0.556 (0.516)
0.001	0.539 (0.534)	0.590 (0.534)	0.570 (0.521)	0.581 (0.532)	0.645 (0.546)	0.712 (0.544)	0.567 (0.549)
0.0001	0.533 (0.518)	0.551 (0.542)	0.568 (0.526)	0.624 (0.542)	0.647 (0.540)	0.716 (0.540)	0.565 (0.543)
0.00001	0.521 (0.525)	0.553 (0.509)	0.526 (0.536)	0.607 (0.558)	0.642 (0.528)	0.717 (0.540)	0.545 (0.515)
Log odds method							
0.8	0.668 (0.529)	0.595 (0.534)	0.610 (0.530)	0.614 (0.541)	0.646 (0.530)	0.721 (0.531)	0.601 (0.518)
0.5	0.668 (0.531)	0.592 (0.531)	0.610 (0.525)	0.618 (0.536)	0.642 (0.534)	0.724 (0.518)	0.601 (0.513)
0.1	0.636 (0.547)	0.580 (0.534)	0.599 (0.523)	0.598 (0.519)	0.652 (0.522)	0.743 (0.515)	0.574 (0.522)
0.05	0.620 (0.537)	0.560 (0.527)	0.596 (0.524)	0.592 (0.521)	0.656 (0.526)	0.747 (0.526)	0.568 (0.523)
0.01	0.567 (0.548)	0.600 (0.509)	0.585 (0.517)	0.574 (0.522)	0.666 (0.530)	0.749 (0.525)	0.569 (0.529)
0.001	0.533 (0.527)	0.590 (0.528)	0.580 (0.519)	0.597 (0.535)	0.661 (0.547)	0.749 (0.545)	0.575 (0.558)
0.0001	0.528 (0.520)	0.545 (0.544)	0.571 (0.534)	0.627 (0.544)	0.658 (0.557)	0.748 (0.534)	0.569 (0.549)
0.00001	0.529 (0.521)	0.556 (0.522)	0.520 (0.531)	0.612 (0.555)	0.655 (0.533)	0.749 (0.533)	0.544 (0.526)

The values in plain font are the median AUC statistics produced when nominally associated SNPs are used to discriminate case–control status for the same disease. The values in parenthesis are median AUC statistics produced when nominally associated bipolar SNPs are used to discriminate case–control status in other diseases (or coronary heart disease SNPs for bipolar cases). BD, bipolar disorder; CHD, coronary heart disease; HT, hypertension; CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type I diabetes; T2D, type II diabetes.

In an ideal scenario, each and every variant contributing to disease susceptibility as well as its associated relative risk would be known a priori. However, for the vast majority of complex disorders, our knowledge of the genetic architecture underlying disease susceptibility is far from complete. For most common diseases, only a few risk predisposing variants of small to moderate effect are known, and together they explain only a small amount of the total phenotypic variation. Indeed, most of the heritability is still unaccounted for even in the case of 'well-characterized' diseases for which upwards of 30 predisposing loci have been identified (13). One possibility is that at least some of this 'missing heritability' is attributable to variants of small effect spread widely across the genome.

In this study, we investigate whether information from GWAS could be harnessed to improve discrimination of complex disease affection status. The current paradigm within genetic diagnostics is to test individuals only at loci known to affect disease risk—yet the technology exists to genotype an individual at hundreds of thousands of loci across the genome. In this regard, large case–control studies that employ the genome-wide association approach contain information on thousands of loci that display nominal levels of significance. Although many of these nominal associations will represent statistical fluctuation and type I error, others will reflect true loci of small effect that do not meet the stringent levels required for statistical significance. We investigate whether this information can be exploited to improve the discriminative ability of genetic testing for complex diseases, if an individual was to be genotyped across their genome. Such an approach has an inherent appeal, since explicit knowledge of the identity of loci that contribute to disease risk is unnecessary.

RESULTS

Table 1 displays median area under the curve (AUC) values for the seven diseases using the count and log odds methods. The values in plain font are the AUC statistics produced when nominally associated SNPs are used to discriminate case–control status for the same disease (i.e. the 'profiling conditions'). The values in parenthesis are the AUC statistics produced in the 'baseline conditions', i.e. when nominally associated bipolar SNPs are used to discriminate case–control status in the other disease groups (or coronary heart disease SNPs in the case of bipolar cases). A full set of results including the range of AUC values produced across the ten prediction sets is presented in Supplementary Material, Table S1.

Median AUC values were relatively low across all conditions for bipolar disorder, coronary heart disease, hypertension, Crohn's disease, rheumatoid arthritis and type II diabetes, suggesting that at least for these diseases, genome-wide scores have little discriminative ability on their own. In contrast, the median AUC scores for the type I diabetes data were moderate, often in excess of 0.7. Repeating the analysis of type I diabetes with SNPs in and around the major histocompatibility complex (MHC) excluded from calculation of the genome-wide score resulted in median AUC values from 0.568 to 0.594, suggesting that SNPs within the MHC were primarily responsible for the increased levels of discriminatory accuracy (data not shown).

Interestingly, in the case of bipolar disorder, coronary heart disease, hypertension and type II diabetes, the genome-wide scores had most discriminative utility when liberal cut-offs were used to select SNPs in the training set (i.e. $\alpha = 0.8$ or $\alpha = 0.5$). Indeed for these diseases, there was a trend for the median AUC to decrease as the threshold for including loci in calculation of the score became more stringent. In contrast, for rheumatoid arthritis and type I diabetes, the best discrimination was achieved using stringent thresholds, although not when the calculations were repeated with MHC SNPs excluded when a liberal threshold resulted in better discrimination (data not shown). We also note that for most diseases

Table 2. Median AUC values for known variants and known variants plus genome-wide scores combined

Threshold	BD	CHD	CD	RA	T1D	T2D
Count method						
Known	0.549	0.572	0.769	0.701	0.784	0.666
0.8	0.657 (0.564)	0.624 (0.579)	0.782 (0.770)	0.716 (0.703)	0.793 (0.784)	0.702 (0.670)
0.5	0.671 (0.566)	0.619 (0.576)	0.780 (0.770)	0.718 (0.704)	0.794 (0.785)	0.670 (0.667)
0.1	0.651 (0.561)	0.593 (0.581)	0.771 (0.770)	0.718 (0.712)	0.787 (0.785)	0.690 (0.667)
0.05	0.656 (0.556)	0.589 (0.580)	0.770 (0.771)	0.715 (0.712)	0.787 (0.785)	0.686 (0.667)
0.01	0.608 (0.584)	0.608 (0.569)	0.770 (0.771)	0.716 (0.708)	0.788 (0.785)	0.669 (0.665)
0.001	0.563 (0.561)	0.597 (0.572)	0.770 (0.770)	0.710 (0.709)	0.786 (0.785)	0.668 (0.665)
0.0001	0.574 (0.561)	0.576 (0.576)	0.771 (0.770)	0.709 (0.709)	0.785 (0.787)	0.669 (0.669)
0.00001	0.561 (0.562)	0.579 (0.578)	0.770 (0.769)	0.703 (0.712)	0.785 (0.786)	0.669 (0.668)
Log odds method						
0.8	0.678 (0.572)	0.618 (0.585)	0.779 (0.770)	0.718 (0.708)	0.792 (0.786)	0.707 (0.668)
0.5	0.674 (0.566)	0.617 (0.580)	0.778 (0.770)	0.719 (0.709)	0.793 (0.786)	0.707 (0.666)
0.1	0.641 (0.562)	0.595 (0.583)	0.772 (0.770)	0.718 (0.715)	0.788 (0.785)	0.696 (0.667)
0.05	0.641 (0.562)	0.594 (0.579)	0.769 (0.771)	0.718 (0.715)	0.788 (0.786)	0.687 (0.667)
0.01	0.597 (0.579)	0.620 (0.573)	0.769 (0.772)	0.713 (0.711)	0.788 (0.785)	0.668 (0.666)
0.001	0.560 (0.563)	0.592 (0.576)	0.769 (0.770)	0.712 (0.714)	0.785 (0.784)	0.669 (0.667)
0.0001	0.569 (0.561)	0.577 (0.573)	0.770 (0.772)	0.710 (0.710)	0.784 (0.790)	0.667 (0.671)
0.00001	0.560 (0.562)	0.577 (0.581)	0.770 (0.770)	0.703 (0.713)	0.787 (0.785)	0.671 (0.673)

The first row displays the AUC achieved by using known variants only to discriminate case–control status. The values in the rows below this show the AUC achieved using known variant information combined with genome-wide scores. The values in plain font are the median AUC statistics produced when known variants plus nominally associated SNPs are used to discriminate case–control status for the same disease. The values in parenthesis are median AUC statistics produced when known variants for the disease of interest are combined with genome-wide scores derived from nominally associated bipolar SNPs (or coronary heart disease SNPs for bipolar cases). BD, bipolar disorder; CHD, coronary heart disease; HT, hypertension; CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type I diabetes; T2D, type II diabetes.

there was little difference between using the count method and the log odds method to generate a genome-wide score, although the log odds method tended to perform better when predicting type I diabetes and rheumatoid arthritis (again when MHC SNPs were excluded from the calculation, the count and log odds method performed similarly—data not shown).

It is also revealing to examine how well genome-wide scores discriminated case–control status in the baseline analyses. Median AUC values for the baseline analyses tended to be low across most of the diseases and conditions (i.e. AUC < 0.55) suggesting that although not a major factor, at least some of the discriminative utility of the genome-wide scores might reflect batch effects, genotyping error and/or population stratification in the control sample. The difference between the profiling and baseline conditions tended to be greatest at liberal thresholds (i.e. $\alpha = 0.8$ or $\alpha = 0.5$) in the case of bipolar disorder (0.13–0.14 difference in AUC), coronary heart disease (0.06 difference in AUC), hypertension (0.07–0.09 difference in AUC) and type 2 diabetes (0.06–0.07 difference in AUC). Examination of the full range of possible AUC values across the ten prediction sets, suggested that liberal thresholds tended to produce a large difference between profiling and baseline conditions, whereas this difference was not apparent when strict thresholds were employed (Supplementary Material, Table S1). In contrast, in the case of Crohn's disease, type I diabetes and rheumatoid arthritis the difference in AUC between the profiling and baseline conditions was most apparent and reliable at stringent thresholds. Although there was little difference in median AUC between profiling and baseline conditions for Crohn's disease, examination of the ranges of possible AUC values suggested that this difference was most reliable for stringent thresholds particularly for the log odds method (Supplementary Material, Table S1).

Supplementary Material, Table S2 displays median odds ratios obtained by comparing individuals in the highest and lowest genetic-score quintiles. The highest median odds ratios for each disease (i.e. according to the best threshold and allele scoring method) varied widely from moderately large (~3–4) for hypertension, type II diabetes and coronary heart disease, to very large for type I diabetes (more than 10). The pattern of odds ratios mirrored the AUC results closely. Similar to the AUC, relaxing the *P*-value threshold tended to increase the median odds ratio for bipolar disorder, coronary heart disease, hypertension, Crohn's disease and type II diabetes, but reduced it for rheumatoid arthritis and type I diabetes.

The top row of Table 2 displays median AUC values generated by genotyping known variants in the prediction set (i.e. the row labelled 'Known'). In general, discrimination was poor for bipolar disorder, coronary heart disease and type II diabetes, but moderate for Crohn's disease, rheumatoid arthritis and type I diabetes where AUC values reached more than 0.7. Comparing median AUC scores between Table 1 and the top row of Table 2 showed that genotyping known variants resulted in better discrimination of affection status than using genome-wide scores in the case of Crohn's disease, rheumatoid arthritis, type I diabetes and type II diabetes, but not bipolar disorder or coronary heart disease.

Table 2 also displays the median AUC values resulting from adding genome-wide scores to known variant information. Including genome-wide information resulted in a substantial gain in median discriminative accuracy relative to just using known variants in the case of bipolar disorder (+0.13 AUC), coronary heart disease (+0.05 AUC) and type 2 diabetes (+0.04 AUC) and was most apparent when liberal thresholds were used to select SNPs. In addition, the small increase in AUC using known variants and genome-wide scores derived from unrelated bipolar SNPs relative to just using known variants, suggested that the increased discriminative ability in the profiling conditions was probably not a result of batch effects, genotyping errors or stratification in the control group. Similar results were observed when considering the full range of AUC values across the ten prediction datasets (Supplementary Material, Table S3). In contrast, the addition of genome-wide information only produced a small increase in median AUC in the case of rheumatoid arthritis, Crohn's disease and type I diabetes relative to using known variant information.

DISCUSSION

Our results indicate that genome-wide scores constructed via the count and log odds methods provide a low to moderate amount of discrimination in affection status, which are currently unlikely to be of diagnostic utility on their own. Nevertheless, this result is interesting because it implies that even if the genetic variants contributing to disease risk are unknown, it is still possible to derive a genetic score that has some discriminative ability using genome-wide information. This fact is highlighted particularly in the case of hypertension where there are no known common risk predisposing loci and yet it is still possible to construct a crude genome-wide measure that provides a median AUC of 0.61. Both genome-wide methods performed best in discriminating affection status in type I diabetes. Type I diabetes is an auto-immune disease that has a substantial MHC conferred susceptibility component. It is likely that the genome-wide scores included a strong contribution from the many strongly associated SNPs in the MHC and this increased their discriminative ability. Repeating the analyses with SNPs from the MHC excluded confirmed this.

For most diseases, it mattered little in terms of discriminative accuracy whether genome-wide scores were constructed using the count method or the log odds procedure. This result is similar to Janssens *et al*. (14) who found little difference in discriminative accuracy when genomic profiling was performed by counting the number of risk genotypes in each profile or by calculating the associated disease risks. Our result implies that for many conditions the differences in effect sizes of individual loci were too minor to affect the discriminative ability of whole genome profiling. The exceptions were type I diabetes and to a lesser extent rheumatoid arthritis, both diseases that have a major MHC contribution. The reason is that at low *P*-value thresholds, genome-wide scores for these diseases primarily reflect genuine risk loci of large effect from the MHC, which provide good discrimination of affection status. In contrast, as the *P*-value threshold becomes less severe, the genome-wide scores become 'contaminated' by unassociated SNPs and markers that reflect a much smaller risk contribution. As a result, the discriminatory power provided by the MHC SNPs becomes attenuated.

Genotyping known variants resulted in median AUC values that ranged from 0.549 in the case of bipolar disorder, to a moderate 0.784 in the case of type I diabetes. Genotyping known variants usually provided superior discrimination of affection status compared with genome-wide scores. This is not surprising since as well as including a small amount of genuine signal from truly associated loci, the genome-wide scores also include noise from hundreds or thousands of other loci scattered throughout the genome that are not associated (or have only very minimal associations) with the disease. In contrast, known variants provide 'clean' information that is uncontaminated by this noise. The exceptions were bipolar disorder and coronary heart disease where the genome-wide scores did a better job of discriminating between cases and controls than the known variants did, but in these diseases only three and one confirmed variant of small effect have been identified.

The inclusion of genome-wide scores in addition to known variant information resulted in a small increase in the ability to discriminate affection status. This result implies that there are SNPs on the genome-wide chip, or variants tagged by them, that are associated with disease and are still awaiting discovery/confirmation. Interestingly, the largest increases in discrimination tended to occur at liberal thresholds, suggesting (consistent with quantitative genetics theory) that there are many loci of small effect located in the lower part of the test statistic distribution. This was most apparent in the case of bipolar disorder where the increase in median AUC produced by adding the genome-wide scores was most pronounced, and to a smaller extent also in coronary heart disease and type II diabetes. In contrast, in the case of Crohn's disease, rheumatoid arthritis and type I diabetes, little was gained from adding genome-wide information. This makes sense intuitively, since many loci have been discovered which influence risk of these conditions. It is likely that most of the common loci with the largest effects and consequently the greatest discriminatory power have been discovered for these conditions, and the addition of genome-wide information from such a small sample was unlikely to have much predictive value. In contrast, only three and one confirmed loci have been discovered for bipolar disorder and coronary heart disease, respectively, allowing for the possibility that many undiscovered loci of moderate effect may still exist.

Although the discrimination of affection status was improved with the addition of genome-wide information, it is important to realise that an increase in discriminative accuracy does not mean that such a result will be potentially useful clinically. In this study, the majority of conditions were associated with small changes in the AUC, which are unlikely to be diagnostically useful, at this stage. However, the results are encouraging that the diagnostic value of genome-wide information will become increasingly useful in the future as larger datasets become available (15). The results of GWAS have demonstrated empirically that the first generation of such studies were not powered, on the whole, to detect the majority of variants with effect size equal to those confirmed in the follow-up replication studies (16). The research community has rallied to merge GWAS samples in mega-analyses. For example, the psychiatric GWAS consortium aims to have more than 13 000 cases and 13 000 controls for each of five psychiatric disorders (17). Such large datasets are expected to increase the number of known validated associated variants.

However, they are also likely to make the genome-wide information more useful since SNPs in the lower end of the *P*-value distribution will more likely reflect truly associated loci (15). With larger samples, we might also expect that the stringency of the threshold for inclusion of SNPs may increase since the distributions of true- and false-positive associations will be pulled apart a little further.

We realise that there is likely to be a degree of bias associated with our results. First, many of the known variants were discovered using the Wellcome Trust Case Control Consortium (WTCCC) dataset and so are likely to fit the prediction set better than they would in independent samples. We consider this to be of minor concern since our focus is on the genome-wide scores, not the discriminative ability of the known variants. Secondly, it is possible that the known variants and the genome-wide scores did not optimally discriminate between affected and unaffected individuals because some of the unselected 1958 British Birth Cohort controls were in fact affected. This issue should be relatively minor since most of these diseases have low prevalence and so the majority of control individuals will be truly unaffected, although this may be a problem for more common conditions such as coronary heart disease and type II diabetes where the discriminative ability of the genetic tests may have been artificially reduced. This lower discrimination between cases and controls represents lower effective power and may contribute to the higher *P*-value thresholds for inclusion of SNPs in prediction sets for these diseases compared with the low prevalence diseases of Crohn's disease and type I diabetes.

We also acknowledge that our genome-wide scores will not capture non-additive relationships (i.e. genetic dominance within loci and epistasis between loci), as well as genetic variation that is not attributable to or tagged by SNPs on the genome-wide chip (i.e. possibly including structural variation and copy number variants). Although, we have used the AUC measure to quantify the discriminative ability of our genome-wide scores (as it is an accepted measure of the efficacy of diagnostic tests), we acknowledge that it is a population value, which may obscure the fact that some individuals at high risk may be able to be identified (15,18). Moreover, its maximum value will depend on the heritability of the disease. For example, it is theoretically possible to obtain excellent discrimination (i.e. maximum AUC > 0.95) for rarer, highly heritable diseases like bipolar disorder, but not for more common diseases where heritability is likely to be less than 50% (e.g. coronary heart disease) in which case only moderate levels of discrimination (e.g. AUC $= 0.8$) may be theoretically possible (12). The AUC values presented in this study are still well below the theoretical maximums reported in Janssens *et al.* (12) and suggest that there exists more heritable information which could be utilized for diagnostic purposes. In addition, the inclusion of other information such as sex, age, and well-known environmental risk factors is likely to improve discriminative ability even further.

Finally, we acknowledge that some of the discrimination afforded by genome-wide scores might be due to genotyping error, batch effects and/or population stratification. We took several steps to exclude the influence of these potential confounders including imposing strict filters on tests of Hardy–Weinberg equilibrium and missingness. We also included a baseline condition where we used SNPs from the bipolar disorder 1958 Birth Cohort comparison to generate a list of nominally associated SNPs to predict case–control status in the other diseases. The fact that these conditions only produced low-AUC values suggests that the discrimination afforded by genome-wide scores was not due to genotyping confounders, at least in the control group. However, we acknowledge that these sources of variation may still be present within the case samples and inflate the discrimination of case–control status afforded by the genome-wide scores. We also cannot rule out the possibility that bipolar SNPs genuinely contribute to the genetic aetiology of the other diseases. If this was the case then we would expect the AUC in the baseline bipolar conditions to be inflated, and we would consequently overestimate the role played by batch effects/genotyping error/population stratification and conversely underestimate the degree to which genome-wide scores discriminate case–control status. Given the low-AUC values for most of the baseline bipolar conditions in Table 1, any genuine effect of bipolar SNPs on the other diseases is likely to be minor. We also note that this possibility does not detract from our principal conclusion—that genome-wide scores increase ability to discriminate between cases and controls.

In conclusion, we have shown how the addition of genome-wide information using the count and log odds methods results in a small increase in discriminating case–control status. We conclude that this small increase in discriminative accuracy is unlikely to be of diagnostic or predictive utility at the present time.

MATERIALS AND METHODS

We employed previously published data from the WTCCC (16). Briefly, the WTCCC is a GWAS involving individuals with one of seven diseases: bipolar disorder (1868 individuals), coronary heart disease (1926 individuals), Crohn's disease (1748 individuals), hypertension (1952 individuals), rheumatoid arthritis (1860 individuals), type I diabetes (1963 individuals) or type II diabetes (1924 individuals), as well as a common set of 1480 unselected controls from the 1958 British Birth Cohort. Individuals were genotyped using the Affymetrix 500K SNP chip. Genotype data were subjected to rigorous quality control measures in order to remove poor quality SNPs as well as putatively related individuals and those of non-European ancestry [for a full description of the cohorts and the data cleaning procedures applied to the data see the original WTCCC article (16)]. In addition, in order to ensure that only the cleanest genotype data contributed to the calculation of the genome-wide scores, we also excluded any SNP in Hardy–Weinberg disequilibrium (exact *P*-value less than 0.05 in cases or controls), SNPs that differed in missing rate between cases and controls ($P < 0.05$), and any SNP with MAF $< 1\%$. These stringent quality control criteria were performed in order to increase the likelihood that genetic differences between cases and controls were due to risk predisposing loci rather than batch effects, genotyping error and/or population stratification.

Each disease sample together with the same set of 1958 British Birth Cohort controls were split into two samples—a

'Training Set' (90% of cases and 90% of controls) and a 'Prediction Set' (the remaining 10% of cases and 10% of controls). Individuals in the Training Set were used to identify a genome-wide set of SNPs that might potentially predispose to disease risk. Armitage trend tests were performed across the genome in the training set using the PLINK program (19). Loci with *P*-values lower than a certain threshold on the Armitage trend test were then selected to discriminate affection status in the Prediction Set. Allelic tests of association were also performed on this subset of loci in the Training Set individuals in order to estimate their allelic odds ratios.

We investigated two simple ways of combining information from genome-wide SNP data. In the first method, the total number of risk alleles an individual carries, both within and across loci was counted:

$$N(\text{risk}) = \sum x_i$$

where x_i=number of risk alleles (=0, 1, 2) at SNP *i* (by 'risk allele' we mean the allele which displays greater frequency in cases than controls in the particular training set). We refer to this method of combining genome-wide information as the 'Count method'. As this method essentially assumes that all risk alleles contribute equally to disease risk, we might expect this method to perform poorly in diseases where there are a mixture of different effect sizes (e.g. auto-immune diseases).

The second method sums together the natural logarithm of the allelic odds ratio for each risk allele within and across loci:

$$\log(\text{risk}) = \sum x_i \times \log(\text{OR}_i)$$

where OR_i is the allelic odds ratio as estimated in the Training Set. We refer to this method as the log odds method. As this procedure incorporates information about effect size, we might expect it to perform better than the count method when risk loci with a mixture of underlying effect sizes contribute to disease. Loci with *P*-values for association that are far from significant, add little to the overall log(risk) score, since $\log(\text{OR}) \approx \text{OR} - 1 \rightarrow 0$.

A key question for both methods is whether there is an optimal threshold for deciding which loci to include in the calculation of the genome-wide score. A threshold that is too liberal is likely to incorporate noise and hence obscure any true signal, whereas a threshold that is too strict, risks discarding loci that genuinely contribute to disease risk. As the effect of the threshold on discriminative ability is likely to vary across different diseases, we investigated the performance of eight different thresholds ranging from very liberal to very strict (i.e. $\alpha = 0.8$; $\alpha = 0.5$; $\alpha = 0.1$; $\alpha = 0.05$; $\alpha = 0.01$; $\alpha = 0.001$; $\alpha = 0.0001$; $\alpha = 0.00001$). At each threshold, we calculated the odds ratio obtained by comparing individuals in the highest and lowest genetic-score quintiles.

In diagnostics, the discriminative ability of a diagnostic test is usually evaluated in regard to two quantities: sensitivity and specificity. Sensitivity is the probability of a positive-test result given the individual examined is truly affected by disease. Specificity is the probability of a negative-test result given the individual is not affected by disease. A perfect diagnostic test has a sensitivity of one (i.e. all individuals who develop the disease have a positive result) and a specificity of one (i.e. all individuals who do not have the disease have a negative result). For composite tests, which involve a number of different components (e.g. a panel of genetic tests), positive- and negative-test results are defined by a cut-off value for the probability of disease. The sensitivity and specificity of the composite test varies depending upon the cutoff probability chosen. It is possible to calculate the sensitivity and specificity for each possible cutoff value and plot the results in a Receiver Operator Characteristic Curve (20). The AUC quantifies the discriminative ability of the diagnostic test and is equivalent to the familiar C-statistic from logistic regression. The AUC ranges from 0.5 indicating a total lack of discrimination to $\text{AUC} = 1$ indicating perfect discrimination. The AUC can be considered the probability of correctly identifying the diseased subject from a pair of subjects, one diseased and the other not. For example, an AUC of 0.95 means that 95% of such pairs are classified correctly, where as an $\text{AUC} = 0.50$, means that only half of pairs are classified correctly—no better than expected by chance (12). As a rough guide, a test with an $\text{AUC} \sim 0.8$ might be useful in screening individuals who are at increased risk of disease, whereas much higher values of AUC are needed to convincingly diagnose a disease before the onset of observable symptoms (12).

We quantified the discriminative accuracy of the genome-wide scores that were generated at each of the eight thresholds in the Prediction Set using the AUC measure. This provided an indication of the potential utility of the genome-wide score, which did not rely on explicit knowledge of the risk variants underlying the conditions. However, the predictive ability of genome-wide scores might primarily reflect the influence of known variants of moderate effect (or loci in LD with them), rather than polygenic loci of smaller effect scattered throughout the genome. We were therefore interested in whether the genome-wide scores could improve diagnostic accuracy over and above testing the known variants for each disease.

We calculated AUC values for each disease using known variants or the best available proxy SNP that was present on the Affymetrix 500K chip (Supplementary Material, Table S4). Logistic regression models were fitted to the data assuming an additive model for each known SNP on the logit scale (i.e. equivalent to a score of 0, 1 or 2 for each SNP). The only disease where this was not possible was hypertension, since no confirmed loci have been associated with common forms of this disease. We then reconstructed the genome-wide scores excluding all known variants as well as SNPs 1 Mb either side of the known variant to ensure there was no contamination from loci in linkage disequilibrium with the known variants. In the case of rheumatoid arthritis and type I diabetes that are autoimmune diseases, we also excluded all SNPs around the MHC on chromosome 6 from 25 to 35 Mb (based on Build 35 positions). We then recalculated the AUC measures. To ensure proper comparability of the discriminative ability of the known variants and the discriminative ability of the genome-wide scores, we only performed analyses in individuals that were genotyped successfully at all known loci (i.e. the loci listed in Supplementary Material, Table S4).

As successful discrimination between case and control samples might reflect batch effects, genotyping error and/or latent population stratification rather than genuinely associated SNPs, we repeated all of our analyses using SNPs that were nominally associated with bipolar disorder to generate genome-wide scores and predict case–control status for the other diseases (in order to predict individuals with bipolar disorder we used coronary heart disease SNPs to generate the genome-wide scores). Since we would expect the majority of SNPs underlying bipolar disease to be different from those underlying risk of the other diseases, any ability to discriminate between the other disease cases and controls using bipolar SNPs (i.e. an AUC > 0.5) is likely to reflect the presence of batch effects, genotyping error and/or population stratification. To facilitate ease of presentation, we refer to these sets of analyses as the baseline analyses. We refer to analyses where nominally associated SNPs are used to discriminate case–control status for the same disease as the profiling conditions.

Finally, in order to get an estimate of the precision of our results, we used a 10-fold cross validation procedure. For each of 10 analyses, the training set consisted of 90% of samples and the prediction set was the remaining 10%. The 10 prediction sets were chosen to be mutually exclusive and of equal size. From the results of these 10 analyses, the median and range of AUC were estimated.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We would like to thank Shaun Purcell, Jonathan Sterne and Tom Palmer for valuable discussions. We thank the Wellcome Trust Case Control Consortium for making the genotype data available.

Conflict of Interest statement. None declared.

FUNDING

D.M.E. is supported by a Medical Research Council New Investigator Award (MRC G0800582). N.R.W. and P.M.V. are supported by Australian National and Medical Research Council grants 389892, 442915, 339450, 443011 and 496688.

REFERENCES

1. Lamberts, S.W. and Uitterlinden, A.G. (2008) Genetic testing in clinical practice. *Annu. Rev. Med.*, **60**, 431–442.
2. Norton, M.E. (2008) Genetic screening and counseling. *Curr. Opin. Obstet. Gynecol.*, **20**, 157–163.
3. Altshuler, D., Daly, M.J. and Lander, E.S. (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.
4. Collins, F.S. and McKusick, V.A. (2001) Implications of the human genome project for medical science. *J. Am. Med. Assoc.*, **285**, 540–544.
5. Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
6. Valle, D. (2004) Genetics, individuality, and medicine in the 21st century. *Am. J. Hum. Genet.*, **74**, 374–381.
7. Holtzman, N.A. and Marteau, T.M. (2000) Will genetics revolutionize medicine? *N. Engl. J. Med.*, **343**, 141–144.
8. Vineis, P., Schulte, P. and McMichael, A.J. (2001) Misconceptions about the use of genetic tests in populations. *Lancet*, **357**, 709–712.
9. Pharoah, P.D., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F. and Ponder, B.A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.*, **31**, 33–36.
10. Yang, Q., Khoury, M.J., Botto, L., Friedman, J.M. and Flanders, W.D. (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am. J. Hum. Genet.*, **72**, 636–649.
11. Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M.J. and Seddon, J.M. (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat. Genet.*, **38**, 1055–1059.
12. Janssens, A.C., Aulchenko, Y.S., Elefante, S., Borsboom, G.J., Steyerberg, E.W. and van Duijn, C.M. (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet. Med.*, **8**, 395–400.
13. Maher, B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.
14. Janssens, A.C., Moonesinghe, R., Yang, Q., Steyerberg, E.W., van Duijn, C.M. and Khoury, M.J. (2007) The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet. Med.*, **9**, 528–535.
15. Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, **17**, 1520–1528.
16. The Wellcome Trust Case Control Consortium (2009) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
17. Sullivan, P.F. (2008) Schizophrenia genetics: the search for a hard lead. *Curr. Opin. Psychiatry*, **21**, 157–160.
18. Wray, N.R., Goddard, M.E. and Visscher, P.M. (2008) Prediction of individual genetic risk of complex disease. *Curr. Opin. Genet. Dev.*, **18**, 257–263.
19. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. and Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
20. Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.