Prediction of individual genetic risk of complex disease
Naomi R Wray¹, Michael E Goddard² and Peter M Visscher¹

Most common diseases are caused by multiple genetic and environmental factors. In the last 2 years, genome-wide association studies (GWAS) have identified polymorphisms that are associated with risk to common disease, but the effect of any one risk allele is typically small. By combining information from many risk variants, will it be possible to predict accurately each individual person’s genetic risk for a disease? In this review we consider the lessons from GWAS and the implications for genetic risk prediction to common disease. We conclude that with larger GWAS sample sizes or by combining studies, accurate prediction of genetic risk will be possible, even if the causal mutations or the mechanisms by which they affect susceptibility are unknown.

Introduction
Common complex diseases, such as psychiatric disorders, cancer, diabetes, heart disease and asthma, are caused by multiple genetic and environmental factors. Significant heritabilities and increased risk to relatives quantify the importance of the genetic factors. To predict the risk of a complex disease for a healthy individual we need to know and be able to measure risk factors, their effect sizes and how they interact. Although prediction of total risk is an ultimate goal, prediction of genetic risk, the risk that can be attributed to inherited genetic variants, is an important component and is the focus of this review.

Predictive genetic tests are already available for a huge range of Mendelian disorders, those for which a single genetic mutation is known to cause the disease [1], but for common complex diseases, very few causal genetic risk factors have been identified. Consequently genetic prediction has been mostly limited to family history information. However, the value of family history information in clinical diagnosis is limited: for example, all children in a nuclear family are predicted to have the same genetic risk based on the history of disease in their parents and more distant relatives, yet they are genetically different. Great progress has been made in the last 2 years in the identification of common polymorphisms that are associated with risk of disease in the population. Now that these risk variants have been identified, can they be used to predict an individual’s genetic risk for a particular disease more accurately than can be done using family history information?

The possible impact of prediction of genetic risk on individual and population health has been recognized for sometime [2,3,4], but it is only the new developments in high-density genotyping technology [5] that make genetic risk prediction within reach. In this review, we discuss the recent advances made by genome-wide association studies in identifying genetic variants associated with disease and the way in which these results can be used for ‘genomic profiling’ [6], the prediction of genetic risk to disease.

Genetic architecture of complex diseases
Genetic architecture of a disease refers to the number of genetic polymorphisms that affect risk of disease, the distribution of their allelic frequencies, the distribution of their effect sizes and their genetic mode of action (additive, dominant and/or epistatic). Prediction of genetic risk is dependent on the underlying genetic architecture because as the number of causal variants increases, the proportion of variance explained by each decreases. As a consequence, it becomes harder to detect them experimentally and to estimate their allelic effects accurately.

Genome-wide association studies (GWAS) – what can they tell us about genetic architecture of complex diseases?
Following the sequencing of the human genome and the creation of a map of common haplotypes [7], commercial panels of ~300 to ~500 thousand single nucleotide polymorphisms (SNPs) were created that covered most common variation. These SNP chips have been used extensively in the last two years in large GWAS for common diseases and quantitative traits. To date, more than 25 (e.g. [8]) of these studies have been published and many more are under way. Validation studies have shown that at least some of the associations found in GWAS are replicated in independent samples. What
Frequency distribution of effect sizes expressed as Odds ratio for the risk allele of 92 validated associated SNPs identified from GWAS. These SNPs represent associations with one of 16 disorders (listed in Appendix A). The power of the GWAS to detect variants with effect size of 1.1 or smaller was low.

Figure 1

general conclusions can we draw from these recent studies about the genetic architecture of complex diseases?

(i) Most associated variants that have been detected are common (minor allele frequency, MAF >0.05), although this is mostly a reflection of study design in terms of SNPs selected for genotyping and power of the sample size. Risk variants are almost equally likely to be the minor or major alleles.

(ii) Although a few variants of large effects (allelic Odds ratio, OR > 2) have been detected, the vast majority of the effect sizes of risk alleles are small, typically OR <1.5 but many around 1.1 and 1.2 (Figure 1), which are the limits of detection given the experimental sample sizes employed to date.

(iii) For most diseases, the detected variants explain little of the total genetic variance that we know exist (from twin and family studies). For breast cancer, prior to GWAS, about 20% of the familial risk could be accounted for by variants in six genes (16% from BRCA1 and BRAC2 alone) [9,10]. Results from GWAS have identified variants that account for only an additional 2.3% of the familial risk [9].

(iv) The genetic mode of action is additive on the logarithm of risk (log-risk) scale (multiplicative on the risk scale) for the majority of variants identified, as expected from the significant heritabilities [11], but the power to detect interactions between variants has been limited by the sample sizes used to date [8**].

Together these studies point to a genetic architecture of a few risk variants of large or moderate size, but a very substantial number of variants which generate small increases in disease risk [8**]. But why has so little of the genetic variance been explained? First, the effect sizes may be underestimated as the associated variants may only be in linkage disequilibrium (LD) with the causal variants. Second, the genotyped markers may not be representative of all the inherited genetic variance, especially structural variation [12,13,14]. Thirdly, the GWAS SNP panels are limited to polymorphisms with MAF >0.01. Rare causal variants are likely to exist but will not be represented (‘tagged’) by the genotyped SNPs, as high r^2 LD between SNPs requires a close matching of allele frequencies [15]. However, such variants cannot individually (by virtue of being rare) explain much of the genetic variance [16,17*].

Since most of the identified associated SNPs have effect size close to the limit dictated by the power of the studies, a likely explanation, at least in part, is that there are many common polymorphisms with effects that were too small to pass the stringent significance thresholds. Given the effect size of variants detected so far, study samples of the order of 10,000 cases and controls would be needed to detect variants which explain the majority of genetic variance [8**,18*].

Prediction of genetic risk from genome-wide markers

Many genetic variants each conferring only a small increased risk to disease are individually not useful in predicting a person’s genetic risk to disease. However, a risk equation combining presence/absence of each risk variant and its effect size can generate a personalised prediction of genetic risk. Given the emerging evidence for the genetic architecture underlying complex diseases, how accurate would a prediction equation be in predicting individual genetic risk? The SNPs (or other markers) do not have to be the causative mutations: they just need to be in high LD with the causative mutations so that there is a consistent association between the SNP and disease risk. We investigated this problem using simulation of GWAS based on realistic assumptions of underlying genetic architecture [18*]. For simplicity we assumed perfect LD between causal and genotyped SNPs and that the genotyped SNPs could explain all the genetic variance by combining additively on the log-risk scale. Our results confirmed that GWAS of only a couple of thousand of cases and controls (the current norm) are still too small to detect most of the causal variants and those detected will not explain enough of the genetic variance to be useful in risk prediction (Figure 2). However, for a case-control study with 10,000 cases and controls our results show that even for diseases controlled by 1000 loci with mean relative risk of only 1.04 (and maximum ~1.3), it will be possible to identify ~75 of the larger loci that explain >50% of the genetic variance and yield a correlation between predicted and true genetic risk >0.7 [18*]. If 500 loci affect disease risk,
the correlation is $\rho = 0.9$. Therefore, as results from multiple GWAS are combined, a larger fraction of the genetic variance is likely to be explained and accurate prediction of genetic risk to disease will become possible even though the risks conveyed by individual variants are small.

A practical issue is identification of optimal statistical methods for selecting and combining SNPs into a prediction equation. The central problem of selecting a set of predictors when the number of measured variables is very large and greater than the number of observed phenotypic values is a common statistical challenge – the large $p$ small $n$ paradigm – arising in a wide range of diverse fields (including gene expression data, for example [19]) and a number of these methods have been applied to SNP selection [20–25,26]. Just as the individual SNP associations discovered in GWAS need to be validated in independent datasets, prediction equations estimated from one dataset need to be validated in other independent data.

Assessing the utility of prediction of genetic risk

The precision of a prediction of genetic risk can be assessed by the correlation ($\rho$) between true and predicted genetic risk or by $\rho^2$ which is the proportion of the genetic variance explained by the associated variants in data independent of that used to identify them. The magnitude of $\rho$ depends on the proportion of the genetic variance tagged by the genotyped markers and the accuracy with which their effects on risk are estimated.

Whereas $\rho$ is a tangible measure for quantitative traits, for disease traits it is perhaps best expressed at the proportion of the sibling risk ratio ($\lambda_s$) explained by the prediction (Figure 2).

The utility of the multi-marker equation for predicting absolute risk of disease depends on the heritability of the trait. If the heritability is low, the value of predicting the genetic component of risk in predicting disease will also be low. The ability to predict disease status given perfect knowledge of genetic risk ($\rho = 1$) was the subject of [27]. The receiver operator characteristic (ROC) curve can be used to compare the sensitivity and specificity of a genomic profile test in correctly classifying diseased and non-diseased individuals. However, the maximum area under the curve explained by a genetic prediction is limited by the heritability of the disease (and also disease prevalence if applied to a population sample). Therefore, we have argued that use of ROC statistics can be confusing for assessing the accuracy of the genetic component of risk prediction [18].

The epidemiological measure population attributable fraction (PAF) can be used to quantify the decrease in disease prevalence if the identified set of risk factors is eliminated. PAF measures the change in the mean of disease prevalence and can be high even when the proportion of genetic variance explained by the markers is low. For example, the 11 validated variants for type 2 diabetes identified by GWAS [28], have risk allele
frequencies ranging from 0.31 to 0.87 and OR ranging from 1.10 to 1.37; together they explain <5% of the genetic variance or $\lambda_s$ but have PAF > 80%. PAF describes the effect of eliminating risk alleles on the population mean whereas $\rho^2$ quantifies the precision of genetic prediction to individuals. These two measures represent utility to the so-called ‘population’ and ‘high-risk’ based approaches to prevention of common disease, discussed in detail in the context of genomic profiling by [6].

The use of genetic risk prediction
Prediction of genetic risk will be useful in diagnosis, treatment, prognosis and prevention strategies, as family history is currently, but with greater power. For diseases with very low population prevalence, genetic risk prediction might be limited to those with family history of the disease, providing differential risk prediction for family members. For higher prevalence disease, genetic risk prediction could be applied at a population level to identify a sub-group genetically most at risk. Figure 3 shows that, even for diseases with $\lambda_s$ of only 2, the majority of future disease cases occur in the 15% of people with highest genetic risk [29]. If we assume that environmental (E) risk factors, like genetic (G) risk factors, act additively on the log-risk scale, then on the risk scale they will act multiplicatively, implying substantial G $\times$ E interaction on this scale. Therefore, the effects of environmental risk factors or intervention strategies will be larger for those individuals with high genetic risk to disease. For example, if an intervention strategy reduces relative risk by 50% then people with a 40% genetic risk of disease can cut their absolute risk to 20% whereas people at a low genetic risk of 1% will cut their risk to 0.5%. Therefore, it would be efficient use of resources to target the intervention to the small proportion of the population at high risk (Figure 3). Such strategies would have even more importance if G $\times$ E is present on the log-risk scale.

A number of studies have already used associated genetic variants to predict genetic risk of disease, often in combination with environmental factors [30–35], but their success has been limited by the small proportion of genetic variance explained by the variants identified to date. Time will tell if larger study samples in GWAS are sufficient for identification of variants that explain the majority of the genetic variance. For some diseases (e.g. psychiatric disorders [36,37]), increasing phenotypic homogeneity of clinical diagnosis in samples used in GWAS may be needed before accurate predictions can be made. Moreover, biologically driven, genotypically defined diagnoses, so-called ‘reverse phenotyping’ [38], may result [37].

Implementation of risk prediction in a clinical context has the potential to be of major economic benefit to population health [4*]. It also raises serious ethical and social concerns [4*,39,40], but these may have been by-passed as November 2007 saw the launch of three new companies (Decodeme, 23andme, Navigenics) offering genomic profile services for less that US $1000.

**Figure 3**

The cumulative proportion of disease cases present in the population when ranked from lowest to highest genetic risk for diseases with different $\lambda_s$ (after [29]). Even for diseases with $\lambda_s$ of only 2, 70% of cases occur within the top 27% and 50% of cases occur in the top 12% of those ranked on genetic risk. These results are independent of disease prevalence, although $\lambda_s$ is constrained to be low for high prevalence diseases.
Conclusion
The results of GWAS provide empirical evidence that the genetic architecture of complex disease is one of many common causal variants each of small, additive effect. GWAS have not yet identified variants that explain enough of the genetic variance to make accurate predictions of genetic risk. However, simulation studies and experiments suggest that with larger sample sizes or by combining studies, accurate prediction of genetic risk will be possible. In the long term, complete sequencing of each person’s genome [41] means that risk of a disease could be predicted from all sequence variants. In the short term, genome variation can be represented by the genotyping of a representative set of SNPs. The value from predicting individual disease risk from multiple associated variants could be reaped long before the causal mechanism of each is determined. Cost-effective intervention strategies could be targeted to the subset of the population most at genetic risk to disease. In the long-term prediction of this additive genetic risk can be combined with predictions from environmental risk factors, non-additive genetic factors and interactions between them, along with other molecular (such as gene or protein expression) profiles [42]. However, an individual’s genomic profile could be available from birth, long before exposure to most environmental risk factors takes place.

Acknowledgements
This work was supported by the Australian National Health and Medical Research Council grants 389892, 442915, 443011 and 496688.

Appendix A
References for the 16 disorders used in Figure 2:

1. Abdominal aortic aneurysm [43].
2. Age-related macular degeneration [44].
3. Amyotrophic lateral sclerosis [45].
4. Ankylosing spondylitis [46].
5. Asthma [47].
6. Autoimmune thyroid disease [46].
7. Breast cancer [48].
8. Childhood obesity [49].
9. Colorectal cancer [50].
10. Coronary heart disease [51,52].
12. Prostate cancer [57,58].
13. Rheumatoid arthritis [59,60].
15. Type 1 diabetes [64,65].
16. Type 2 diabetes [28].

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:  
- of special interest  
- of outstanding interest

8. WTCCC: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007, 447:661-678.
9. The pioneering GWAS from the Wellcome Trust Case Control Consortium of Bipolar Disorder, Coronary Artery Disease, Crohn’s Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes and Type 2 Diabetes.
We use simulation and a quantitative genetic framework to investigate the value of genomic profiling.


43. Discussion about the use of molecular profiling (combining of any molecular measures for use in disease diagnosis, prevention or treatment) in clinical decision making are relevant also to genomic profiling. Nat Genet 2008, 40:217-224.


