GENOME-WIDE ASSOCIATION STUDIES — OPINION

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

Abstract | The success of genome-wide association studies (GWASs) has led to increasing interest in making predictions of complex trait phenotypes, including disease, from genotype data. Rigorous assessment of the value of predictors is crucial before implementation. Here we discuss some of the limitations and pitfalls of prediction analysis and show how naive implementations can lead to severe bias and misinterpretation of results.

In many species, single-nucleotide polymorphism (SNP)–trait associations have been detected through genome-wide association studies (GWASs). In addition to the discovery of trait-associated variants and their biological function, there is increasing interest in making predictions of complex trait phenotypes from genotype data for individuals in plant and animal breeding, experimental organisms and human populations. These predictions are based on selections of SNPs (or other genomic variants) and estimation of their effects in a discovery sample, followed by validation in an independent sample with known phenotypes and ultimately application to samples with unknown phenotypes (FIG. 1).

The validation stage of SNP prediction analysis will be the main focus of this Perspective. Incorrect conclusions at this stage may lead to predictors that will not work as well as inferred or in the worst case will have no prediction accuracy at all. We organize our Perspective into limitations and common pitfalls of prediction analysis. The limitations are partly inherent, given the nature of the trait or the data available. These are factors that users should be aware of but mostly cannot change. The limitations also reflect use of suboptimal methodology that could be improved on. The pitfalls are common mistakes in analysis that can lead to overestimation of the accuracy of a predictor or misinterpretation of results, and we give examples from the literature where these have occurred. We give our opinion on how best to avoid pitfalls in the derivation and application of SNP-based predictors for practical applications. There are many aspects of risk prediction that are outside the scope of this article. They include a thorough treatment of the statistical methods that can be used in the discovery phase[1–7], the use of non-genetic sources of information to make predictions or diagnosis, a full discussion about clinical utility of risk prediction in human medicine and a discussion about ethical considerations for applications in human populations[8].

Limitations of prediction analyses

Limitation 1: prediction of phenotypes from genetic markers. Variation in complex traits is almost invariably due to a combination of genetic and environmental factors. A useful quantification of the importance of genetic factors is the heritability (h^2): that is, the proportion of phenotypic variation in a trait that is due to genetic factors[9] (BOX 1). Assuming that the estimated h^2 is a true reflection of the population parameter, then h^2 is the upper limit of the phenotypic variance explained by a linear predictor (R^2) based on DNA markers such as SNPs, and a genetic predictor can thus never fully account for all phenotypic variation. This upper limit is achievable only if all genetic variants that affect the trait are known and if their effects are estimated without error. In human disease genetics, in which 'personalized medicine' is actively being pursued, this limitation is not well understood in our opinion, and hence we have chosen to highlight it here, even though it has been pointed out before[10,11].

Environmental risk factors can be added to the genetic predictor to make a better predictor of the phenotype. In practice, not all environmental factors are identified (and some factors that are classified as 'environment' may simply be stochastic events[12]). For example, combining SNPs and phenotypic predictors — such as body mass index and smoking — improved prediction of age-related macular degeneration, which is an eye disease in humans in which age is a major risk factor[13]. In some circumstances, more accurate phenotyping, including the use of repeated measures, can lead to a more heritable trait. In general, expectations need to be adjusted accordingly for the application of phenotype or disease prediction in humans.

Unlike the deterministic genetic tests for fully penetrant Mendelian disorders, genetic predictions for complex traits will be probabilistic, and the value may only be incremental in clinical decision making. The value of genetic risk prediction may be at a group level rather than an individual level. For example, from a risk predictor for type 1 diabetes (T1D), which was created from risk variants known up to 2011, a risk group comprising the top ranked 18% of individuals would need to be monitored to capture 80% of future cases. However, because T1D is not common (it has a prevalence of 0.4%), the probability of disease for individuals in this risk group is still less than 2%[14]. Nonetheless, cost-effective public health strategies could result from use of genetic predictors to identify high-risk strata in which disease prevention interventions should be focused[15,16]. In agriculture, genetic risk prediction is mostly geared towards selection of breeding stock on the basis of estimates of additive genetic values (that is, estimated breeding values) in the parent generation; the aim is to elicit average changes in the phenotype of the offspring generation. That is, the impact of genetic prediction is naturally at the level of a group rather than an individual.

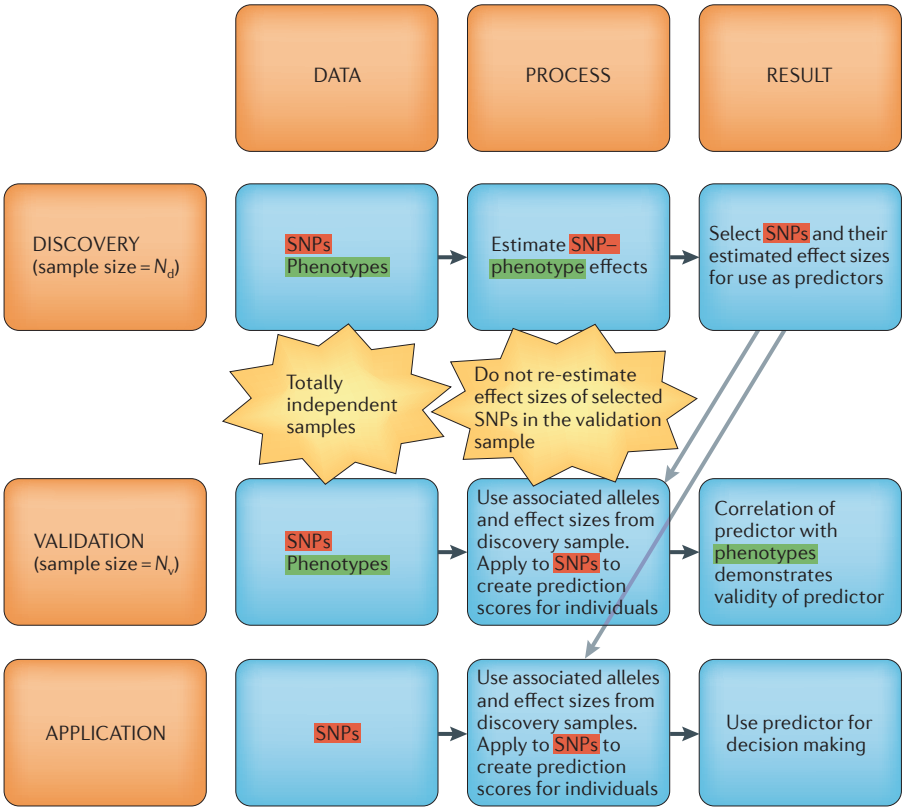


Figure 1 | **Flowchart of SNP-based prediction analysis.** There are three stages for the development of a risk predictor: discovery, validation and application. At each stage, data are needed as an input, and a process is applied to the data and a result is generated. At the application stage, effect sizes estimated from combined discovery and validation samples can be used. SNP, single-nucleotide polymorphism.

Limitation 2: variance explainable by markers. The SNPs included in the genome-wide SNP chips used for identifying SNPs associated with complex traits are typically not the causal variants for a phenotype — it is more likely that they may have an association with the trait because they are in linkage disequilibrium (LD) with one or more causal variants. As the SNPs on SNP chips are chosen because both of their alleles are common, they cannot be in complete LD with a causal variant with one rare allele. If the variation generated by the causal variants is completely explained by the genotyped SNPs, then the SNPs can potentially explain all of the genetic variation in the trait (that is, $h^2_M = h^2$, where h^2_M is defined as the genetic variation captured by the SNPs or markers). Sometimes (see, for example, REF. 17), h^2_M is referred to as the 'narrow-sense heritability'; however, in our opinion, the term narrow-sense heritability should be reserved as the definition of the total additive genetic variance: that is, h^2 (REFS 9,18).

If a genetic variant is associated with fitness, selection will drive one allele to low frequency[19–21]. This is the case even for traits without an obvious connection to fitness. The larger the effect of a SNP on fitness, the lower the frequencies of the causal alleles are expected to be[22,23]. For example, individual mutations that cause severe intellectual disability in humans are rare[24,25]. Therefore, in practice, the SNPs identified as associated in the discovery population are unlikely to explain all genetic variation (that is, $h^2_M < h^2$) as contributions to the variance by rare variants may not be tagged by the genotyped SNPs[26–28]. For example, for both height and schizophrenia, $h^2 \approx 0.7$–0.8 and $h^2_M \approx 0.5$ for height[26] and $h^2_M \approx 0.2$–0.3 for schizophrenia[29,30].

The difference between the variance explained by genome-wide-significant SNPs (h_{GWS}) and heritability estimates from family studies (h^2) has been called the 'missing heritability', and the difference between h_{GWS} and h^2_M has been described as the 'hidden heritability'. As such, the difference between h^2_M is referred to as the 'still missing heritability': that is, $h_{GWS} < h^2_M < h^2$. The still missing heritability may simply reflect genomic variants that are not well tagged by SNPs. In livestock populations, when missing heritability is defined in this way, little is missing, and up to 97% of the heritability is captured by common SNPs[31,32], probably because the smaller effective population size leads to long-range LD, and hence even rare alleles can be predicted by a linear combination of SNPs that are in LD with the causal variant. Even in dairy cattle, however, traits that could reasonably be assumed to be under strong natural selection, such as fertility, have greater missing heritability[31]. Moreover, when the SNPs are fitted together with a pedigree, as much as half of the genetic variance is explained by the pedigree and not by the SNPs[33]. The simplest explanation is that in livestock as in humans, some causal variants are rare and in poor LD with the SNPs.

With the advances in whole-genome sequencing technologies, causative mutations will be present in the data, and the proportion of variation that can be captured by the sequence data is expected to approach h^2. In principle, known rare risk variants, if identified, can be included in the predictor in the same way as common variants; cumulatively, their contribution may be important. Their importance can be assessed by the proportion of variation that they explain. Both the ability to detect an association between a trait and a SNP and the value of including the SNP in a predictor depend on the proportion of variance the SNP explains. For example, a rare variant with a frequency of 1 in 1,000 in the population and a relative risk for a disease of five will increase the risk of disease by fivefold for 1 in 1,000 people (so, from 1% to 5% for a disease with a prevalence of 1%), but such an increase in risk can also be achieved by the cumulative effect of multiple common variants with smaller effect size. The contribution of rare variants can be included in a predictor by grouping them into defined classes of genes[34,35] or by incorporating prior knowledge of functions[36].

Limitation 3: errors in the estimated effects of the markers. The effects of SNPs on a trait must be estimated from a sample of finite size, and so the effects are estimated with some sampling error. If there were only a few loci that affected a trait, it would be possible to estimate their effects quite accurately, but most complex traits are controlled by a large number of largely unknown loci[37].

Therefore, the discovery stage of estimating the prediction equation may involve a genome-wide panel of millions of SNPs. The true effects of most SNPs are small, and so the accuracy with which these effects are estimated is low unless a large discovery sample is used. The correlation between a phenotype and a predictor that uses all SNPs simultaneously in a randomly mating population can be expressed as a function of effective population size (or the effective number of independent chromosome segments, which is a function of effective population size), heritability and the size of the discovery sample[38–40] (see equation 1 in BOX 1). Specifically, SNP effects will be better estimated when the sample size of the discovery cohort increases (see the figure in BOX 1); estimated or predicted effect sizes of rare variants will be difficult to verify even with large sample sizes.

Limitation 4: statistical methods in the discovery sample. The least squares prediction or 'profile scoring'[29] method is commonly used for prediction of genetic risk. Although it is simple to apply, it does not have desirable statistical properties, and an arbitrary *P* value threshold is used for the selection of SNPs that go in the predictor. Moreover, estimating SNP effects one at a time is not an optimal approach[1,41–46]. This is because SNP effects are correlated, and accounting for LD in the profile scoring method requires SNP selection on arbitrary thresholds. Methods that model the distribution of SNP effects[42] and the correlation between SNPs in the presence of single as well as multiple causal variants will be more accurate[1,41–45,47]. In human applications, sometimes only genome-wide-significant SNPs are included in the predictor[15,48–51], yet greater accuracy results from the use of less stringent thresholds[1,39,42], and in animal and plant breeding, it is typical to use all available SNPs. Better SNP estimation methods exist and are used in plant and animal breeding[1,2,39,46,52]; such methods have been proposed for applications to human data[1,45]. They rely on prior assumptions about the distribution of SNP effects in the genome and use all data simultaneously. Such Bayesian methods have also been applied to other species[53], and related methodologies derived in computer science have been applied to disease data in humans[4,54]. Ignorance cannot be bliss in this context, and it must be best to use all available genetic and phenotypic information simultaneously. It is outside the scope of this Perspective to discuss these methods in more detail.

Box 1 | **Quantifying phenotypic variation explained by SNPs**

Quantitative traits

The proportion of phenotypic variance (R^2) explained by a predictor of a quantitative trait formed using estimated effects of all markers depends on the number (M) of independently measured genomic variants (for example, single-nucleotide polymorphisms (SNPs)) associated with the trait, the proportion of the total variance they explain (h^2_M) and the sample size in the discovery sample (N_d)[27,38,40]. If all marker effects are assumed to come from the same normal distribution, then approximately

$$R^2 = \frac{h^2_M}{1 + \frac{M}{N_d h^2_M}(1 - R^2)}. \quad (1)$$

Equation 1 holds regardless of the genetic architecture of the trait, but we note that the predictor may be far from optimal. h^2_M is usually less than the heritability estimated from family studies and is sometimes called the SNP heritability or chip heritability and is estimated, for example, using GCTA[63]. Equation 1 is taken from the supplement of REF. 40; when R^2 is small, it can be ignored in the denominator, otherwise the quadratic in R^2 can be solved. The graph shows that the sample size must be large in order to achieve a high R^2. If the distribution of marker effects sizes is markedly non-normal, with some large effects and many very small or zero effects, and if knowledge of this distribution is used in estimating SNP effects, then higher R^2 can be achieved[65].

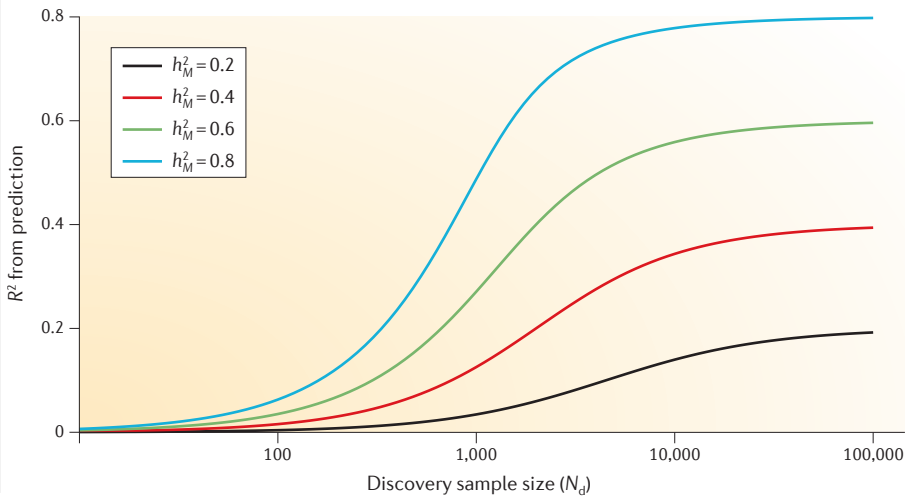
In this article, we use R^2 as the statistic to report efficacy of a predictor or R, the correlation between phenotype and predictor or accuracy. The sign of the correlation is important for interpretation of the predictor. In livestock, genetic predictors have been used for decades (on the basis of pedigree data prior to the availability of genotypic data), and accuracy ($R_{G,\hat{G}}$) is traditionally used to evaluate utility. $R_{G,\hat{G}}$ is the correlation between true and estimated genetic value (the predictor, which is an estimate of the combined value of all genetic loci). Because

$$R^2_{G,\hat{G}} = \frac{R^2}{h^2} \quad (2)$$

the $R_{G,\hat{G}}$ statistic quantifies the efficacy of a genetic predictor relative to the best possible genetic predictor.

Disease traits

For disease traits, Nagelkerke's R^2 (R^2_N) has been used in profile scoring analyses, following Purcell *et al.*[29]. R^2_N is an R^2 measure in binary (0–1) outcome data and is usually applied in case–control validation samples, in which the proportion of cases is much higher than in the population. Alternatively, the area under the receiver operator curve (AUC) is reported[73–75]: this is a statistic with a long tradition of use in determining the efficacy of clinical predictors. AUC has the desirable property of being independent of the proportion of cases in the validation sample; one definition of AUC is that a randomly selected case is ranked higher by the predictor than a randomly selected control. A new statistic reflecting variance explained on the liability scale (R^2_l), which can be related to other statistics, such as R^2_N and AUC[11], has been proposed[76]. Like any estimate on the liability scale, calculation of R^2_l requires specification of disease prevalence in the population but allows direct comparison of the variance explained by the predictor to estimates of heritability from family data and estimates of SNP heritability from genome-wide SNP data.



Pitfalls of the analysis

Pitfall 1: validation and discovery sample overlap. If the correlation (R) between a phenotype and a single SNP in the population is zero (that is, if the SNP is not associated with the trait), the expected value of the squared correlation (R^2) estimated from a sample of size N is $1/(N–1)$ or approximately $1/N$ if N is large. Hence, a randomly chosen 'candidate' (but not truly associated) SNP explains $1/N$ of variation in any sample. Usually, $1/N$ is small enough not to worry about. However, a set of m uncorrelated SNPs that have nothing to do with a phenotype of interest would, when fitted together, explain m/N of variation (owing to summing of their effects). For example, when fitted together in a regression analysis in a discovery sample of $N_d = 1{,}000$, a set of 100 independent SNPs would, on average, explain 10% ($R^2 = 0.10$) of phenotypic variance in the discovery sample under the null hypothesis of no true association. (Note that the effective number of independent markers (M) from standard GWAS chips is estimated to be ~60,000 in European populations, a number that is based on analyses of LD[29], genomic inflation factors[55] and eigenvalues[56] from principal components analysis. Predictions from theory also come to this number[38].)

When the number of SNPs in the predictor is large and the sample size is small, the discovery R^2 can be very high by chance and can be a gross overestimation of the true variance explained by the predictor when applied in an independent sample. Also, the expected R^2 in the validation sample is $\sim 1/N_v$, where N_v is the validation sample size, for a set of SNPs selected from a discovery sample but with the effect sizes of the SNPs re-estimated in the validation sample. Therefore, to estimate the R^2 of a prediction in a new sample, a prediction equation is estimated in the discovery sample and is tested, without re-estimating the regression coefficients, in the validation sample (BOX 2). Applying the incorrect validation procedure results in over-estimation of the accuracy of the prediction (or over-fitting). An example of a situation in which over-fitting occurs is when testing the prediction in the discovery sample: that is, the same data are used to estimate the effect of SNPs on a phenotype and to make predictions[57,58]. We demonstrate the overlap pitfall with examples in dairy cattle, *Drosophila melanogaster* and human populations (FIG. 2a–c). For example, in a GWAS on ~150 sequenced inbred lines of *D. melanogaster*[58] in which this was done, the authors concluded that 6–10 SNPs selected from >1 million SNPs together explained 51–72% of variation in the lines (depending on the trait analysed). However, a cross-validated Bayesian prediction analysis using all genetic markers on the same data

Box 2 | **Quantifying prediction accuracy for pitfall 2**

When discovery and validation samples are independent

When m single-nucleotide polymorphisms (SNPs) have been selected from a discovery sample, a simple linear predictor in the validation sample is

$$\hat{y} = \sum_{i=1}^{m} \hat{b}_i x_i \quad (3)$$

where $x_i = 0$, 1 or 2 reference alleles of a SNP, and $\hat{b}_i$ is the estimated effect size from the discovery sample. In this article, we do not concern ourselves with how $\hat{b}_i$ is estimated; there are simple least square predictors and more complex Bayesian estimation methods that have been described elsewhere[1,43,44]. We also restrict ourselves to linear (additive) models. Given a multi-SNP predictor ($\hat{y}$), the validation step is to quantify how much of the variation in trait y is explained by the predictor $\hat{y}$. A regression of y on $\hat{y}$ fits only a single covariate so that the R^2 expected by chance is only $1/N_v$, where N_v is the validation sample size. If the validation sample is drawn from the same population as the discovery sample, then a value of $R^2 > 1/N_v$ is evidence for real predictive ability of the predictor. (Software tools output an adjusted R^2 that corrects for the R^2 expected by chance.) Hence the sample size in the validation stage does not have to be large to reject the null hypothesis of no association H_0: $\rho^2 = 0$, where ρ^2 is the true value of R^2 in the population. The standard error (SE) of R is approximately

$$1/\sqrt{N_v} \quad (4)$$

if ρ is very small, and more generally

$$(1-\rho)^2/\sqrt{N_v}. \quad (5)$$

In terms of R^2, its SE is approximately

$$\sqrt{2/N_v} \quad (6)$$

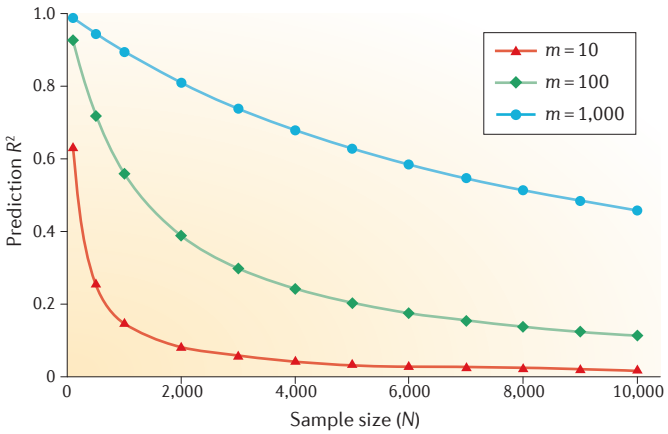
if ρ is small. A general and a more complicated exact equation was given by Wishart[76]. Using the exact equations, if ρ^2 is 1% or 10%, then SE(R^2) for $N_v = 100$ is 1.9% or 5.6% and for $N_v = 500$ it is 0.8% and 2.5%.

When discovery and validation samples are the same

In Supplementary information S1 (box), we derive an approximation of R^2 (verified by simulation) when there is no correlation in the population between SNPs and phenotypes but when m 'associated' SNPs are identified from the same sample (of size N) in which they are validated as a predictor. The relationship between R^2 and N, which is dependent on m and assumes $M = 100{,}000$ independent genomic variants associated with the phenotype, is plotted in the figure, in which m SNPs ($m = 10$, 100 or 1,000) are selected after association analysis of $M = 100{,}000$ uncorrelated SNPs in a sample of unrelated individuals and applied as a predictor back into the same sample when there is no correlation between SNPs and phenotypes. In genome-wide association studies, M is large, so overestimation of R^2 occurs even for big sample sizes.

When validation sample overlaps with the discovery sample

If some of the samples in the validation cohort are also in the discovery set, then this can create spurious results. For the samples that overlap, the expected R^2 between predictor and outcome is the same as in the entire discovery sample, because those samples are just a random sample from the discovery cohort. If the proportion of samples in the discovery set that are also in the validation cohort is q, then the expected squared correlation between predictor and outcome in the entire validation cohort is approximately $q*R^2 + (1-q)/N_v$, with R^2 the (spurious) accuracy derived in Supplementary information S1 (box; see previous section). The important result is that if samples overlap, it is not the proportion of those samples in the discovery cohort that matters, but it is the proportion of the validation samples that is also in the discovery cohort that determines false accuracy.



found that only 6% of phenotypic variation could be explained by the predictor[53].

A less obvious mistake is to select the most significantly associated SNPs in the entire sample and to use these to estimate SNP effects and to test their prediction accuracy in the discovery and validation sets[59]. In this case, the variance explained by the SNPs when applied in the validation sample is inflated. It creates bias and misleading results because the initial selection step of the SNPs is based on there being a chance correlation between these SNPs and the entire sample so also between the SNPs and any subsample. A prediction equation based on these SNPs will appear to work in the validation sample but not in a genuinely independent sample. Cross-validation analysis after the initial set of SNPs has been selected from the entire sample does not mitigate this bias. The pitfall of SNP selection from

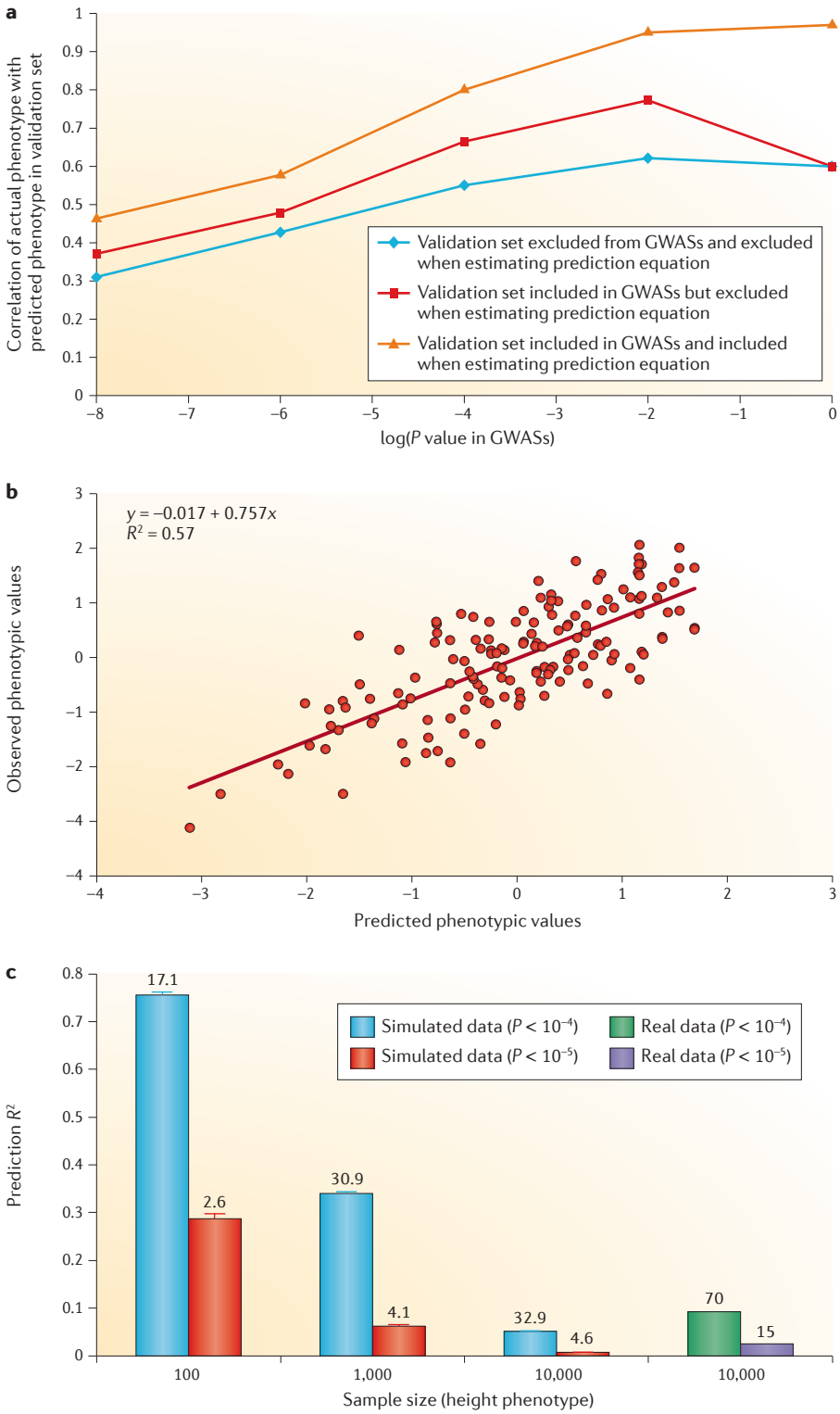


Figure 2 | **The overlap pitfall: non-independence of discovery and validation samples.** **a** | The impact of leaving the validation cohort in the discovery set, either at both single-nucleotide polymorphism (SNP) selection (that is, genome-wide association studies (GWASs)) and SNP effect estimation stages or only at the effect size estimation stage. Data shown are from 2,732 bulls with ~500,000 SNPs phenotyped for average milk yield of their daughters' milk production. The bulls were split into a discovery sample (bulls born during or before 2003), $N_d = 2{,}458$, and a validation sample (bulls born after 2003) of $N_v = 274$. Further methods are included in Supplementary information S3 (box). **b** | An example illustrating bias when selecting the top SNPs. We downloaded genotype data of the *Drosophila* Genetic Reference Panel and simulated phenotypes under the null hypothesis: that is, random association between each of the >1 million SNPs and phenotype. We repeated the GWAS analysis reported in REF. 58, selecting the top 10 independently associated SNPs, and predicted the phenotypes of the lines using these 10 SNPs. Because in the simulated data there are only random associations between a SNP and a phenotype, any prediction power is false and thus a result of over-fitting. By chance, the top SNPs (in terms of test statistic) explain 57% ($R^2 = 0.57$) of the phenotypic variance between the inbred lines, from a linear regression of phenotype on predictor. Both phenotype and predictor have been standardized to normal distribution *Z* scores (with mean of 0 and standard deviation of 1). Further methods are included in the supplementary data. **c** | High R^2 can be achieved by chance, particularly when sample size is small. We simulated GWAS data on the basis of real human genotype data under the null hypothesis of no association representing the phenotype of human height. We used data from 11,586 unrelated European Americans genotyped on 563,212 SNPs[79–81]. We randomly sampled *N* individuals and selected top SNPs for height at $P < 10^{-5}$ (red bar) and $P < 10^{-4}$ (blue bar) to predict the phenotype in the same data. We also carried out association analysis for real height phenotype in 10,000 individuals and selected top SNPs at $P < 10^{-5}$ (purple bar) and $P < 10^{-4}$ (green bar) to predict height phenotype in the same sample. The graph shows the mean prediction R^2 over 100 simulation replicates. The error bar represents the standard error of the mean. The number on top of each column is the mean number of selected SNPs over 100 simulation replicates.

discovery and validation samples occurred in a recent study that reported a genetic predictor of autism[60]. SNPs that were putatively associated with autism in multiple biological pathways were selected on the basis of *P* values from GWASs in the entire data set. Model selection was subsequently applied using cross-validation to narrow down the number of SNPs. The authors did follow up with an independent validation sample, and the prediction accuracy was reduced.

A variation of this pitfall is when a proportion of individuals in the validation sample is also in the discovery sample; the bias is then proportional to the fraction of the validation samples that was also in the discovery set (BOX 2). In practice, it might be difficult to ascertain whether any of the validation individuals were also in the discovery set, in particular if there are only summary statistics available (that is, estimates and standard error of SNP effect and allele frequencies), particularly from public databases. We use cattle data[46] to demonstrate the inflation in variance explained by a SNP predictor when the validation sample is included in discovery steps (FIG. 2c).

The remedy to this pitfall is to use external validation. In some cases, independent data sets are not available, and then internal cross-validation is the only option. In cross-validation, it is important to avoid the pitfall of updating the predictor on the basis of results derived from the validation sample, hence losing the independence of discovery and validation samples that the strategy has set out to achieve[61]. Overlap in samples can be checked as a part of quality control of the prediction pipeline, by estimating pairwise relatedness using SNP data, but this requires access to full genotype data from both discovery and validation samples. There are many software tools that can do this, including PLINK[62] and GCTA[63].

Pitfall 2: the validation sample. If the validation sample is more closely related to the discovery population than to the target population, then the prediction accuracy will be overestimated. In humans, a polygenic prediction analysis of height in 5,117 individuals from the Framingham Heart Study (FHS; original and offspring cohorts only) reported a prediction R^2 of 0.25 using tenfold cross-validation when including all individuals in the analysis[64]. However, because FHS includes many related individuals, the authors repeated the analysis on the basis of pedigree information by restricting the tenfold cross-validation samples to individuals with no known close relatives (that is, parent–offspring, sibling or half sibling) in the data set. In this restricted analysis, the prediction R^2 decreased to 0.15. We caution that cryptic relatedness can still inflate prediction accuracy, even when known close relatives are excluded. To demonstrate this, we conducted a polygenic prediction analysis of height using 7,434 individuals from the FHS SHARe data[65] (BOX 3). Our results demonstrate that cryptic relatedness, beyond the close relatives inferred from pedigrees, can inflate prediction accuracy relative to the prediction accuracy that could be achieved in an independent validation sample.

The remedy of the pitfall described here is to use conventionally unrelated individuals (in discovery and validation stages). Relatedness can be estimated from SNP data[62,63], and so close relatives can be excluded on the basis of observed data. More generally, the validation population should be representative of the population in which the predictor will be ultimately applied. In populations with small effective population size, such as some breeds of livestock, all individuals are related. This does not invalidate the prediction, but it does mean that the same prediction accuracy cannot be expected when the prediction equation is applied to another population that is less closely related to the discovery population[66].

Sometimes, the validation population differs from the application (target) population in that it is much more genetically diverse. For example, the validation (and possibly discovery) population might include a diverse set of lines of animals or plants. A prediction equation may work well in this population but less well in an application population that is less diverse, such as in commercial strains of a crop[66].

Pitfall 3: population stratification similarity. Another way in which prediction accuracy can be inflated is if the discovery and validation samples contain similar patterns of population stratification and if the eventual target population is not similarly stratified. For example, this could occur if discovery and validation samples were independently sampled from a stratified population, such as European Americans[67]. The question of whether this inflation should be viewed as a pitfall depends on the ultimate goal of the analysis. If the goal is to conduct prediction in European Americans,

Glossary

Ancestry principal components
Principal components derived from the genome relationship matrix that account for the genetic substructure of the data. In case–control studies, these principal components can reflect genotyping artefacts, such as plate, batch and genotyping centre, that could be confounded with case–control status.

Conventionally unrelated
Individuals that are not closely related: for example, more distantly related than third cousins.

Cross-validated
Cross-validation involves testing the validity of a prediction in the absence of an independent external validation sample. This is done by dividing the sample into *k* independent subsets (balanced with respect to case–control status in disease data). Each of the *k* subsets is used in turn as a validation sample for a predictor derived from the remaining $k-1$ subsets.

Cryptic relatedness
When a sample is thought to comprise unrelated individuals on the basis of recorded pedigree relationships but in fact includes close relatives: for example, second cousin or closer.

Effective population size
The number of individuals in an idealized population with random mating and no selection that would lead to the same rate of inbreeding as observed in the real population.

Estimated breeding values
Estimates of the additive genetic value for a particular trait that an individual will pass on to descendants.

Heritability
The proportion of phenotypic variance attributable to additive genetic variation.

Independent sample
In the context of risk prediction, this is a sample from the same population but excluding individuals that are closely related. It is necessary for the individuals in different samples from the same population to share common ancestors, and indeed this distant sharing underpins the efficacy of a risk predictor.

Independent SNPs
Uncorrelated single-nucleotide polymorphisms (SNPs) in linkage equilibrium.

Linkage disequilibrium
(LD). The nonrandom association of alleles at different loci.

Polygenic prediction analysis
Any analysis method that predicts genetic risk or breeding values on the basis of the combined contribution of many loci.

Profile scoring
A polygenic prediction method for prediction of genetic value or risk for each individual (a 'profile') in a validation sample generated from the sum of the alleles they carry weighted by the association effect size estimated in a discovery sample.

it is entirely appropriate to leverage ancestry information to the fullest extent possible, and this inflation is not a pitfall (because discovery, validation and target samples are similarly stratified). However, if the goal is to assess the prediction accuracy that could be achieved using less structured application populations, then this inflation is a pitfall. As an example, we show that population stratification was inflating prediction accuracy in the FHS analysis (see BOX 3 for details). A more serious problem is when there is confounding between ancestry and disease status in both discovery and validation case–control samples, because such spurious association can lead to a predictor of ancestry rather than to one of disease. It was recently suggested that the aforementioned predictor of autism[60] suffers from this pitfall[68].

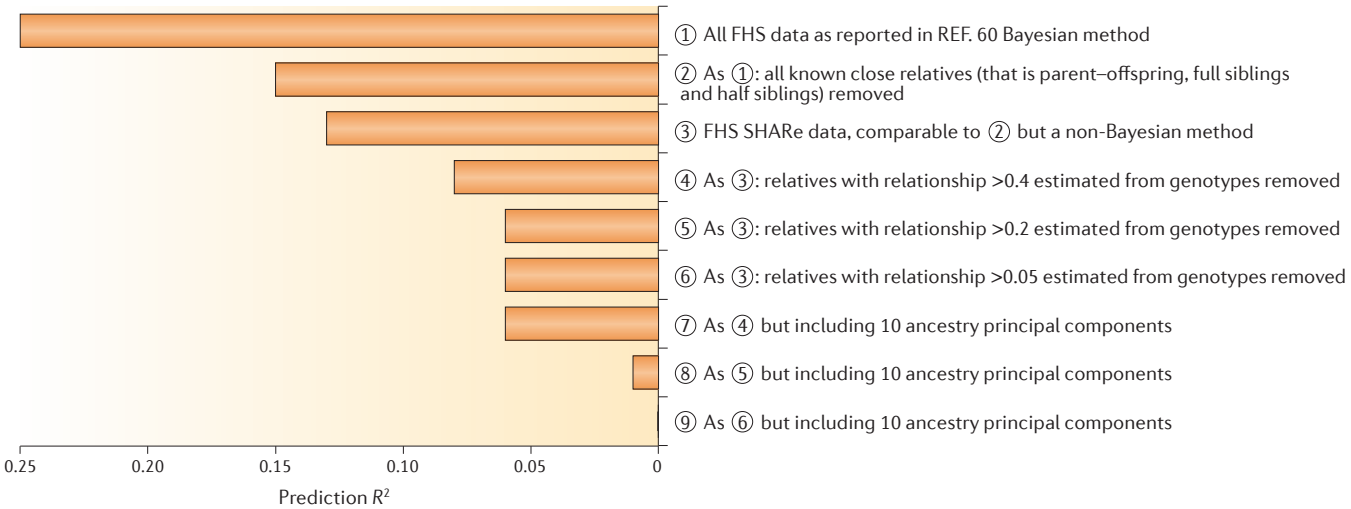
A practical remedy to problems associated with population stratification is to fit ancestry principal components in the analysis of discovery samples. We note that differential bias between cases and controls[69] can also lead to spurious prediction R^2 if discovery and validation samples exhibit the same differential bias as could occur when using tenfold cross-validation. A remedy for differential bias is to carry out stringent quality control and/or to validate predictors in a completely independent sample in lieu of tenfold cross-validation. One quality-control step that can be done is to use the genotyped SNPs that are in the predictor and to quantify the estimated relatedness between the application sample and the discovery and validation samples: for example, in a principal component analysis (PCA)[70] or related methods[71]. If the application sample is an outlier on the PCA, then the prediction accuracy in the target may be less than expected from the validation procedure.

Pitfall 4: expectation of equality of $\mathbf{R^2}$ ***and*** $\mathbf{h^2_M}$. Sometimes called the SNP or chip heritability, an unbiased estimate of the variance explained by markers h^2_M is achieved by correlating phenotypic similarity between pairs of individuals with their SNP-based genotypic similarity[26,63,69]. In human populations, the SNP heritability is broadly between one-third and one-half of total heritability for traits studied to date[28,37,72]. A prediction of phenotype based on the same set of SNPs would achieve $R^2 = h^2_M$ only if the individual SNP effects were estimated without error[27]. For example, when a multiple-SNP predictor that used the 'profile scoring' method was used for height[65], it achieved an R^2 of 0.1–0.15 in out-of-sample predictions. However, Yang *et al.*[26] estimated that all of the SNPs together would explain 40–50% of phenotypic variance if their effects were estimated without error. These results are consistent when the

Box 3 | Using the Framingham Heart Study to demonstrate pitfalls of validation

The Framingham Heart Study (FHS) is a large cohort study of individuals and their family members measured for a wide range of traits (particularly related to cardiovascular disease) and with genome-wide genotypes. A polygenic prediction analysis of height[64] showed that including known related individuals in the analysis inflated R^2 (from 0.15 to 0.25). To investigate whether genetic relatedness can still inflate prediction accuracy even when known close relatives are excluded, we conducted a polygenic prediction analysis of height using 7,434 individuals from the FHS SHARe data[65]. We obtained a prediction R^2 of 0.13 using tenfold cross-validation when restricting to individuals with no known close relatives in the data set on the basis of known pedigree information. (We fit markers individually, whereas in the original study[64], markers were simultaneously fitted by a Bayesian random effects model; thus, it was expected that a slightly higher R^2 of 0.15 was reported.) We repeated the analysis, restricting it to individuals with pairwise relatedness estimated from the single-nucleotide polymorphisms (SNPs) of less than 0.40, 0.20 or 0.05 and obtained prediction R^2 of 0.08, 0.06 and 0.06, respectively, demonstrating the importance of using the genotype data to identify relatives rather than accepting recorded family relationships.

We investigated whether population stratification was inflating prediction accuracy in our FHS analysis, as the prediction R^2 of 0.06 was much higher than would be expected from theory[38] or from empirical data on much larger sample sizes[65]. When repeating the analysis using a height phenotype that was adjusted for 10 eigenvectors[70] of the SNP derived relationship matrix, once again restricting to individuals with pairwise relatedness less than 0.40, 0.20 or 0.05, we obtained prediction R^2 of 0.06, 0.01 and 0.00, respectively, which were smaller values than the prediction R^2 obtained using unadjusted height. The bulk of the reduction came from correcting for the top eigenvector, representing northwest European versus southeast European ancestry[67], which is strongly correlated to height (R^2 = 0.05 in FHS data, which is consistent with other studies[77,78]). Thus, consistent with theory, polygenic prediction analyses of a few thousand unrelated individuals that do not benefit from population stratification will attain a low prediction R^2 (<0.01). The results of these analyses are summarized in the graph. For further details, see Supplementary information S2 (table).



error associated with the estimate of each SNP effect is appreciated.

With ever-larger sample sizes, the size of the error terms in the SNP effect estimates will be reduced, and the two statistics will converge to the same value. However, simulations for human populations suggest that the improvement in trait prediction as sample size increases depends on the genetic architecture of the trait, in particular how many variants there are with tiny effect sizes and that for most common complex genetic diseases the improvement will be slow and modest even when common SNPs account for a large proportion of heritability of the traits[17]. Hence, for applications in human populations to achieve meaningful and accurate predictions, big data are key, and sample sizes of hundreds of thousands needed. Such data sets are starting to become achievable.

Conclusions

We have highlighted what we believe are limitations to genetic risk prediction as well as the most important pitfalls to befall researchers, and we have discussed how these can be avoided. Most problems occur in the validation stage, when data are not fully independent from those in the discovery phase, but care is also needed to ensure that the discovery and validation samples are representative of the population in which the predictor will be applied. Genomic prediction is already having a major impact on livestock selection programmes[39] and has great potential for applications in plant breeding, preventive medicine strategies and clinical decision making. However, there are fundamental limitations to the predictive ability of a genetic predictor (see limitations 1 and 2), and so it is important that expectations are realistic and that the accuracy of genetic predictors is fairly evaluated. As sample sizes increase, predictors of genetic risk will have a greater clinical utility, particularly in terms of identification of population strata at increased risk of disease, as opposed to accurate predictive diagnosis for individuals.

Naomi R. Wray, Jian Yang and Peter M. Visscher are at The Queensland Brain Institute, The University of Queensland, QBI Building, St Lucia, Queensland 4071, Australia.

Jian Yang and Peter M. Visscher are at The University of Queensland Diamantina Institute, Level 7, 37 Kent Street, Translational Research Institute, Woolloongabba, Queensland 4102, Australia.

Ben J. Hayes and Mike E. Goddard are at the Biosciences Research Division, Department of Primary Industries, GPO Box 4440, Melbourne, Victoria 3001, Australia.

Ben J. Hayes is at the Dairy Futures Cooperative Research Centre, AgriBio, Centre for AgriBioscience, 5 Ring Road, La Trobe University, Bundoora, Victoria 3083, Australia; and La Trobe University, Bundoora, Victoria 3086, Australia.

Alkes L. Price is at the Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA; the Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, Massachusetts 02115, USA; the Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; and the Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, Building 2, 2nd Floor, Boston, Massachusetts 02115, USA.

Mike E. Goddard is at the Faculty of Land and Food Resources, University of Melbourne, Melbourne, Victoria 3010, Australia.

Correspondence to P.M.V.
e-mail: peter.visscher@uq.edu.au

doi:10.1038/nrg3457

1. de los Campos, G., Gianola, D. & Allison, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Rev. Genet.* **11**, 880–886 (2010).
2. Gonzalez-Camacho, J. M. *et al.* Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* **125**, 759–771 (2012).
3. Crossa, J. *et al.* Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724 (2010).
4. Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* **5**, e1000678 (2009).
5. de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. L. Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 1255–1268 (2012).
6. Heffner, E. L., Sorrells, M. E. & Jannink, J. L. Genomic selection for crop improvement. *Crop Sci.* **49**, 1–12 (2009).
7. Riedelsheimer, C. *et al.* Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genet.* **44**, 217–220 (2012).
8. Becker, F. *et al.* Genetic testing and common disorders in a public health framework: how to assess relevance and possibilities. *Eur. J. Hum. Genet.* **19**, S6–S44 (2011).
9. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
10. Janssens, A. C. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet. Med.* **8**, 395–400 (2006).
11. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
12. Burga, A., Casanueva, M. O. & Lehner, B. Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* **480**, 250–253 (2011).
13. Seddon, J. M. *et al.* Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest. Ophthalmol. Vis. Sci.* **50**, 2044–2053 (2009).
14. Polychronakos, C. & Li, Q. Understanding type 1 diabetes through genetics: advances and prospects. *Nature Rev. Genet.* **12**, 781–792 (2011).
15. So, H. C., Kwan, J. S., Cherny, S. S. & Sham, P. C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
16. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
17. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genet.* **45**, 400–405 (2013).
18. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nature Rev. Genet.* **14**, 139–149 (2013).
19. Ayodo, G. *et al.* Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am. J. Hum. Genet.* **81**, 234–242 (2007).
20. Raj, T. *et al.* Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. *Am. J. Hum. Genet.* **90**, 720–726 (2012).
21. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
22. Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nature Genet.* **40**, 340–345 (2008).
23. Crow, J. F. Maintaining evolvability. *J. Genet.* **87**, 349–353 (2008).
24. Vissers, L. E. *et al.* A *de novo* paradigm for mental retardation. *Nature Genet.* **42**, 1109–1112 (2010).
25. de Brouwer, A. P. *et al.* Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium. *Hum. Mutat.* **28**, 207–208 (2007).
26. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
27. Visscher, P. M. *et al.* A commentary on 'Common SNPs explain a large proportion of the heritability for human height' by Yang *et al.* (2010). *Twin. Res. Hum. Genet.* **13**, 517–524 (2010).
28. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
29. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
30. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genet.* **44**, 247–250 (2012).
31. Haile-Mariam, M., Nieuwhof, G. J., Beard, K. T., Konstatinov, K. V. & Hayes, B. J. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. *J. Anim. Breed. Genet.* **130**, 20–31 (2013).
32. Jensen, J., Su, G. & Madsen, P. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. *BMC Genet.* **13**, 44 (2012).
33. Kemper, K. E., Daetwyler, H. D., Visscher, P. M. & Goddard, M. E. Comparing linkage and association analyses in sheep points to a better way of doing GWAS. *Genet. Res.* **94**, 191–203 (2012).
34. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
35. Bacanu, S. A., Nelson, M. R. & Whittaker, J. C. Comparison of statistical tests for association between rare variants and binary traits. *PLoS ONE* **7**, e42530 (2012).
36. Lindor, N. M. *et al.* A review of a multifactorial probability-based model for classification of *BRCA1* and *BRCA2* variants of uncertain significance (VUS). *Hum. Mutat.* **33**, 8–21 (2012).
37. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genet.* **44**, 483–489 (2012).
38. Goddard, M. E. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257 (2009).
39. Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**, 433–443 (2009).
40. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008).
41. de los Campos, G. *et al.* Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**, 375–385 (2009).
42. Goddard, M. E., Wray, N. R., Verbyla, K. L. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Statist. Sci.* **24**, 517–529 (2009).

43. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–690 (2009).
44. Guan, Y. T. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Statist.* **5**, 1780–1815 (2011).
45. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
46. Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**, 4114–4129 (2012).
47. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genet.* **44**, 369–375 (2012).
48. Meigs, J. B. *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).
49. Kraft, P. & Hunter, D. J. Genetic risk prediction—are we there yet? *N. Engl. J. Med.* **360**, 1701–1703 (2009).
50. Paynter, N. P. *et al.* Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA* **303**, 631–637 (2010).
51. Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**, 986–993 (2010).
52. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
53. Ober, U. *et al.* Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1002685 (2012).
54. Abraham, G., Kowalczyk, A., Zobel, J. & Inouye, M. SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinformatics* **13**, 88 (2012).
55. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
56. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
57. Derringer, J. *et al.* Predicting sensation seeking from dopamine genes. A candidate-system approach. *Psychol. Sci.* **21**, 1282–1290 (2010).
58. Mackay, T. F. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173–178 (2012).
59. Powell, J. E. & Zietsch, B. P. Predicting sensation seeking from dopamine genes: use and misuse of genetic prediction. *Psychol. Sci.* **22**, 413–415 (2011).
60. Skafidas, E. *et al.* Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Mol. Psychiatry* 11 Sep 2012 (doi:10.1038/mp.2012.126).
61. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002).
62. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
63. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
64. Makowsky, R. *et al.* Beyond missing heritability: prediction of complex traits. *PLoS Genet.* **7**, e1002051 (2011).
65. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
66. Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G. & Hickey, J. M. Genomic prediction in animals and plants: simulation of data, validation, reporting and benchmarking. *Genetics* **193**, 347–365 (2012).
67. Price, A. L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4**, e236 (2008).
68. Belgard, T. G., Jankovic, I., Lowe, J. K. & Geschwind, D. H. Population structure confounds autism genetic classifier. *Mol. Psychiatry* 2 Apr 2013 (doi:10.1038/mp.2013.34).
69. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
70. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
71. Thornton, T. *et al.* Estimating kinship in admixed populations. *Am. J. Hum. Genet.* **91**, 122–138 (2012).
72. Lubke, G. H. *et al.* Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol. Psychiatry* **72**, 707–709 (2012).
73. Machiela, M. J. *et al.* Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epidemiol.* **35**, 506–514 (2011).
74. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **18**, 3525–3531 (2009).
75. Peterson, R. E. *et al.* Genetic risk sum score comprised of common polygenic variation is associated with body mass index. *Hum. Genet.* **129**, 221–230 (2011).
76. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
77. Campbell, C. D. *et al.* Demonstrating stratification in a European American population. *Nature Genet.* **37**, 868–872 (2005).
78. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genet.* **44**, 1015–1019 (2012).
79. Psaty, B. M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.* **2**, 73–80 (2009).
80. Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum. Mol. Genet.* **19**, 2706–2715 (2010).
81. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genet.* **43**, 519–525 (2011).

Acknowledgements
For a full list of acknowledgements, please see the Web version of the article.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Alkes L. Price's homepage: http://www.hsph.harvard.edu/alkes-price
Complex Traits Genomics Group: http://www.complextraitgenomics.com
GCTA Homepage: http://www.complextraitgenomics.com/software/gcta
***Drosophila* Genetic Reference Panel:** http://dgrp.gnets.ncsu.edu
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (table) | S3 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF