Published in final edited form as:
Nat Rev Genet. 2013 July ; 14(7): 507–515. doi:10.1038/nrg3457.

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray[1], **Jian Yang**[1,2], **Ben J. Hayes**[3,4,5], **Alkes L. Price**[6,7,8,9], **Mike E. Goddard**[3,10], and **Peter M. Visscher**[1,2,*]

[1]The University of Queensland, Queensland Brain Institute, Brisbane, Australia

[2]The University of Queensland, Diamantina Institute, Brisbane, Australia

[3]Biosciences Research Division, Department of Primary Industries Victoria, Melbourne, Victoria, Australia

[4]Dairy Futures Cooperative Research Centre, VIC 3083, Australia

[5]La Trobe University, Bundoora, VIC 3086, Australia

[6]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

[7]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America

[8]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

[9]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

[10]Faculty of Land and Food Resources, University of Melbourne, Melbourne, Victoria, Australia

Abstract

The success of genome-wide association studies 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 critical before implementation. Here we discuss some of the limitations and pitfalls of prediction analysis and show how naïve implementations can lead to severe bias and misinterpretation of results.

Introduction

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

Corresponding author: peter.visscher@uq.edu.au.

Web link:
Web applications for the equations in Box 2 are given at http://www.complextraitgenomics.com/software http://pngu.mgh.harvard.edu/~purcell/plink/ http://www.complextraitgenomics.com/software/gcta/

animal breeding, experimental organisms and human populations. These predictions are based upon 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, have no prediction accuracy at all. We organise 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 sub-optimal methodology that could be improved upon. The pitfalls are common mistakes in analysis that can lead to over-estimation 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), i.e. 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, the upper limit of the phenotypic variance explained by a linear predictor (R^2) based on DNA markers such as SNPs is h^2 and a genetic predictor can thus never fully account for all phenotypic variation. This upper limit is only achievable if all genetic variants affecting the trait are known and their effects are estimated without error. In human disease genetics, where ‘personalised 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 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, an eye disease in humans where 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 individual level. For example, from a risk predictor for type 1 diabetes (T1D), created from risk variants known up to 2011, a risk group comprising the top ranked 18% of individuals would need to be monitored in order to capture 80% of future cases, yet because T1D is not common (prevalence 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 where disease prevention interventions should be focussed[15, 16]. In agriculture, genetic risk prediction is geared mostly towards selection of breeding stock based on estimates of additive genetic values (‘**estimated breeding values’**) in the parent generation with the aim of eliciting changes in the phenotype of the of the offspring generation on average. That is, the impact of genetic prediction is naturally at the level of a group rather than an individual.

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 – more likely they may have an association with the trait because they are in **linkage disequilibrium** (LD) with one or more causal variants. Since the SNPs on SNP chips are chosen because both 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 potentially can explain all the genetic variation in the trait (i.e. $h_M^2 = h^2$, where h_M^2 is defined as the genetic variation captured by the SNPs, or markers). Sometimes (e.g.[17]) h_M^2 is referred to as "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 (see 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 a fitness the lower the frequencies of the causal alleles are expected to be[22, 23]. For example, individual mutations causing 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 (i.e, $h_M^2 < h^2$) since 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 ~ 0.7–0.8 and h_M^2 ~ 0.5 for height[26] and 0.2–0.3 for schizophrenia[29, 30].

The difference between the variance explained by genome-wide significant (GWS) SNPs (h_{GWS}^2) and heritability estimate from family studies (h^2) has been called the “missing heritability” and the difference between h_{GWS}^2 and h_M^2 the “hidden” heritability, so that the difference between h_M^2 is the “still missing heritability”, i.e., $h_{GWS}^2 < h_M^2 < h^2$. The still missing heritability may simply reflect genomic variants not well tagged by SNPs. In livestock populations, when missing heritability is defined in this way, little is missing with up to 97% of the heritability 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 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 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 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 1/1000 in the population and a relative risk for a disease of 5 will increase the risk of disease by 5-fold for 1 in 1000 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 into a predictor by grouping them into defined classes of genes[31, 32], or by incorporating prior knowledge of functions[34].

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 very large number of largely unknown loci[35]. 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 very large discovery sample is used. The correlation between phenotype and a predictor that uses all SNPs simultaneously in a random 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 (Equation 1, BOX 1)[36–38]. Specifically, SNP effects will be better estimated when the sample size of the discovery cohort increases (Figure BOX1); 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 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, estimation of SNP effects one at a time is not an optimal approach[1, 39–44]. 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[40] and the correlation between SNPs in the presence of single as well as multiple causal variants will be more accurate[1, 39–43, 45]. In human applications, sometimes only genome-wide significant SNPs are included in the predictor[15, 46–49], yet greater accuracy results from the use of less stringent thresholds[1, 37, 40] 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, 37, 44, 50] and such methods have been proposed for applications to human data[1, 43]. 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[51], and related methodologies derived in computer science have been applied to disease data in humans[4, 52]. Ignorance can't 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.

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, 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 $1/(N\text{-}1)$, 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 (due to the summing of their effects). For example, a set of 100 independent SNPs when fitted together in a regression analysis in a discovery sample of $N_d = 1000$ would, on average, explain R^2 =10% of phenotypic variance in the discovery sample under the null hypothesis of no true association.

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 over-estimation of the true variance explained by the predictor when applied in an **independent sample**. Also, the expected R^2 in the validation sample for a set of SNPs selected from a discovery sample but with the effect sizes of the SNPs re-estimated in the validation sample is $\sim 1/N_v$, with N_v the validation sample size. 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 where over-fitting occurs is when testing the prediction in the discovery sample, i.e., the same data are used to estimate the effect of SNPs on phenotype and to make predictions[53, 54] . We illustrate the overlap pitfall with examples in dairy cattle, *Drosophila* and human populations (FIG 2a-c). . For example, in a GWAS on ~150 sequenced inbred lines of *Drosophila*[54] in which this was done the authors concluded that 6–10 SNPs selected from > 1M 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 found that only 6% of phenotypic variation could be explained by the predictor[51].

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 test their prediction accuracy in the discovery and validation sets[55]. 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 upon there being a chance correlation between these SNPs and the entire sample, so also between the SNPs and any sub-sample. 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 discovery and validation samples occurred in a recent study reporting a genetic predictor of autism[56]. SNPs putatively associated with autism in multiple biological pathways were selected based upon p-values from GWAS 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 on this pitfall is when a proportion of individuals in the validation sample are also in the discovery sample and then the bias is proportional to the fraction of the validation samples that was also in the discovery set (see BOX 2). In practice it might be difficult to ascertain if any of the validation individuals were also in the discovery set, in particular if there are only summary statistics (i.e., estimates and standard error of SNP effect and allele frequencies) available, particularly from public databases. We use cattle data[44] to illustrate 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 in which case internal cross-validation is the only option. In cross-validation it is important to avoid the pitfall of updating the predictor based on results derived from the validation sample, hence losing the independence of discovery and validation samples that the strategy has set out to achieve[57]. Overlap in samples can be checked as part of quality control (QC) 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[58] and GCTA[59].

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 over-estimated. 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 10-fold cross-validation when including all individuals in the analysis[60]. However, because FHS includes many related individuals, the authors repeated the analysis restricting the 10-fold cross-validation samples to individuals with no known close relatives (parent-offspring, sibling, or half-sib) in the data set based on pedigree information. 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[61] (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[58, 59] and so close relatives can be excluded based upon observed data. More generally, the validation population should be representative of the population in which the predictor will ultimately be 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[62].

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 commercial strains of a crop[62].

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 the eventual target population is not similarly stratified. For example, this could occur if discovery and validation samples are independently sampled from a stratified population such as European Americans[63]. 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, 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). On the other hand, 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 both in discovery and validation case-control samples, because such spurious association can lead to a predictor of ancestry rather than one of disease. It was recently suggested that the aforementioned predictor of autism[56] suffers from this pitfall[64].

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[65] can also lead to spurious prediction R^2 if discovery and validation samples exhibit the same differential bias, as could occur when using 10-fold cross-validation. A remedy for differential bias is to perform stringent quality control and/or to validate in a completely independent sample, in lieu of 10-fold cross-validation. One QC step that can be done is to use the genotyped SNPs that are in the predictor and quantify the estimated relatedness between the application sample and the discovery and validation samples, for example in a principal component analysis (PCA)[66] or related methods[67]. 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 R^2 and h_M^2

Sometimes called the SNP- or chip-heritability, an unbiased estimate of the variance explained by markers h_M^2 is achieved by correlating phenotypic similarity between pairs of individuals with their SNP-based genotypic similarity[26, 59, 65]. In human populations, the SNP-heritability is broadly between one-third and one-half of total heritability for traits studied to date[28, 35, 68]. A prediction of phenotype based upon the same set of SNPs would achieve an $R^2 = h_M^2$ 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[61], it achieved an R^2 of 10–15% in out-of-sample predictions. Yet Yang *et al* (2010)[26] estimated that all the SNPs together would explain 40–50% of phenotypic variance if their effects were estimated without error. These results are not inconsistent when the 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 and such data sets are starting to become achievable.

Conclusions

We highlighted what we believe are limitations to genetic risk prediction as well as the most important pitfalls to befall researchers and discussed how these can be avoided. Most problems occur in the validation stage, when data are not fully independent to 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 in livestock selection programmes[37] and has great potential for applications in plant breeding, preventative 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 are fairly evaluated. As sample sizes increase, predictors of genetic risk will have greater clinical utility, particularly in terms of identification of population strata at increased risk of disease as opposed to accurate predictive diagnosis for individuals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge funding from the Australian National Health and Medical Research Council (1047956, 1011506, 613601, 613602, 1048853, 1052684, 1050218), Australian Research Council (FT0991360, DP130102666) and the National Institutes of Health (R01 HG006399, R01 GM 075091, P01 GM 099568, R01 MH100141). We thank John Witte for helpful comments.

Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The human subjects participating in the GWAS derive from The Nurses' Health Study (NHS) and Health Professionals' Follow-up Study (HPFS) and these studies are supported by National Institutes of Health grants CA87969, CA55075, and DK58845. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01HG004424). The datasets used for the analyses described in this manuscript were obtained from dbGap accession number phs000091.

The Atherosclerosis Risk in Communities Study (ARIC) is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. We are grateful to S. Pollack, C. Palmer and J. Hirschhorn for assistance with FHS data.

References

1. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010; 11:880–886. [PubMed: 21045869]
2. Gonzalez-Camacho JM, et al. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics. 2012; 125:759–771. [PubMed: 22566067]
3. Crossa J, et al. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics. 2010; 186:713-U406. [PubMed: 20813882]
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. 2009; 5:e1000678. [PubMed: 19816555]
5. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics. 2012 Published online June 28 2012.
6. Heffner EL, Sorrells ME, Jannink JL. Genomic selection for crop improvement. Crop Science. 2009; 49:1–12.
7. Riedelsheimer C, et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012; 44:217–220. [PubMed: 22246502]
8. Becker F, et al. Genetic testing and common disorders in a public health framework: how to assess relevance and possibilities Background Document to the ESHG recommendations on genetic testing

and common disorders. European Journal of Human Genetics. 2011; 19:S6–S44. [PubMed: 21412252]

9. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008; 9:255–266. [PubMed: 18319743]

10. Janssens AC, et al. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006; 8:395–400. [PubMed: 16845271]

11. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010; 6:e1000864. [PubMed: 20195508]

12. Burga A, Casanueva MO, Lehner B. Predicting mutation outcome from early stochastic variation in genetic interaction partners. Nature. 2011; 480:250–253. [PubMed: 22158248]

13. Seddon JM, 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. 2009; 50:2044–2053. [PubMed: 19117936]

14. Polychronakos C, Li Q. Understanding type 1 diabetes through genetics: advances and prospects. Nat Rev Genet. 2011; 12:781–792. [PubMed: 22005987]

15. So HC, Kwan JS, Cherny SS, Sham PC. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet. 2011; 88:548–565. [PubMed: 21529750]

16. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008; 358:2796–2803. [PubMed: 18579814]

17. Chatterjee N, et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013

18. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet. 2013; 14:139–149. [PubMed: 23329114]

19. Ayodo G, et al. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet. 2007; 81:234–242. [PubMed: 17668374]

20. Raj T, et al. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. Am J Hum Genet. 2012; 90:720–726. [PubMed: 22482808]

21. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]

22. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008; 40:340–345. [PubMed: 18246066]

23. Crow JF. Maintaining evolvability. J Genet. 2008; 87:349–353. [PubMed: 19147924]

24. Vissers LE, et al. A de novo paradigm for mental retardation. Nat Genet. 2010; 42:1109–1112. [PubMed: 21076407]

25. de Brouwer AP, et al. Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium. Hum Mutat. 2007; 28:207–208. [PubMed: 17221867]

26. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]

27. Visscher PM, Yang J, Goddard ME. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al (2010). Twin. Res. Hum. Genet. 2010; 13:517–524. [PubMed: 21142928]

28. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]

29. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]

30. Lee SH, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012; 44:247–250. [PubMed: 22344220]

31. Haile-Mariam M, Nieuwhof GJ, Beard KT, Konstatinov KV, Hayes BJ. 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. 2013; 130:20–31. [PubMed: 23317062]

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. 2012; 13:44. [PubMed: 22694746]

33. Kemper KE, Daetwyler HD, Visscher PM, Goddard ME. Comparing linkage and association analyses in sheep points to a better way of doing GWAS. Genet Res (Camb). 2012; 94:191–203. [PubMed: 22950900]

34. Lindor NM, et al. A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum Mutat. 2012; 33:8–21. [PubMed: 21990134]

35. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012; 44:483–489. [PubMed: 22446960]

36. Goddard ME. Genomic Selection: predicion of accuracy and maximisation of long term response. Genetica. 2009; 136:245–257. [PubMed: 18704696]

37. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009; 92:433–443. [PubMed: 19164653]

38. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS One. 2008; 3

39. de los Campos G, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009; 182:375–385. [PubMed: 19293140]

40. Goddard ME, Wray NR, Verbyla KL, Visscher PM. Estimating effects and making predictions from genome-wide marker data. Statistical Science. 2009; 24:517–529.

41. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009; 10:681–690. [PubMed: 19763151]

42. Guan YT, Stephens M. Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. Annals of Applied Statistics. 2011; 5:1780–1815.

43. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013; 9:e1003264. [PubMed: 23408905]

44. 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. 2012; 95:4114–4129. [PubMed: 22720968]

45. Yang J, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012; 44:369–375. S1-3. [PubMed: 22426310]

46. Meigs JB, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008; 359:2208–2219. [PubMed: 19020323]

47. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? N Engl J Med. 2009; 360:1701–1703. [PubMed: 19369656]

48. Paynter NP, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010; 303:631–637. [PubMed: 20159871]

49. Wacholder S, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010; 362:986–993. [PubMed: 20237344]

50. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819–1829. [PubMed: 11290733]

51. Ober U, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 2012; 8:e1002685. [PubMed: 22570636]

52. Abraham G, Kowalczyk A, Zobel J, Inouye M. SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. BMC Bioinformatics. 2012; 13:88. [PubMed: 22574887]

53. Derringer J, et al. Predicting sensation seeking from dopamine genes. A candidate-system approach. Psychol Sci. 2010; 21:1282–1290. [PubMed: 20732903]

54. Mackay TF, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012; 482:173–178. [PubMed: 22318601]

55. Powell JE, Zietsch BP. Predicting sensation seeking from dopamine genes: use and misuse of genetic prediction. Psychol Sci. 2011; 22:413–415. [PubMed: 21270448]

56. Skafidas E, et al. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Molecular Psychiatry. 2012 epub.

57. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002; 99:6562–6566. [PubMed: 11983868]

58. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

59. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]

60. Makowsky R, et al. Beyond missing heritability: prediction of complex traits. PLoS Genet. 2011; 7:e1002051. [PubMed: 21552331]

61. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]

62. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting and Benchmarking. Genetics. 2012 Published on line Dec 12 2012.

63. Price AL, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 2008; 4:e236. [PubMed: 18208327]

64. Belgard TG, Jankovic I, Lowe JK, Geschwind DH. Population structure confounds autism genetic classifier. Mol Psychiatry. 2013

65. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88:294–305. [PubMed: 21376301]

66. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

67. Thornton T, et al. Estimating kinship in admixed populations. Am J Hum Genet. 2012; 91:122–138. [PubMed: 22748210]

68. Lubke GH, et al. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. Biol Psychiatry. 2012; 72:707–709. [PubMed: 22520966]

69. Yang J, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011; 19:807–812. [PubMed: 21407268]

70. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]

71. Psaty BM, 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. 2009; 2:73–80. [PubMed: 20031568]

72. Qi L, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet. 2010; 19:2706–2715. [PubMed: 20418489]

73. Yang J, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43:519–525. [PubMed: 21552263]

74. Machiela MJ, et al. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol. 2011; 35:506–514. [PubMed: 21618606]

75. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009; 18:3525–3531. [PubMed: 19553258]

76. Peterson RE, et al. Genetic risk sum score comprised of common polygenic variation is associated with body mass index. Hum Genet. 2011; 129:221–230. [PubMed: 21104096]

77. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2542. [PubMed: 22843982]

78. Campbell CD, et al. Demonstrating stratification in a European American population. Nat Genet. 2005; 37:868–872. [PubMed: 16041375]

79. Turchin MC, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet. 2012; 44:1015–1019. [PubMed: 22902787]

80. Gilmour, AR.; Gohel, BJ.; Cullis, BR.; Thompson, R. ASReml User Guide Release 2.0. Hemel Hempstead, UK: VSN International; 2006.

81. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008; 91:4414–4423. [PubMed: 18946147]

82. Wishart MA. The mean and second moment coefficient of the multiple correlation coefficient, in samples from a normal population. Biometrika. 1931; 22:353–361.

Glossary

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

Estimated Breeding Value An estimate of the additive genetic value for a particular trait that an individual will pass on to descendants.

Linkage Disequilibrium The non-random association of alleles at different loci.

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.

Polygenic prediction analysis Any analysis method that predicts genetic risk or breeding values based on 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.

Independent SNPs Independent, uncorrelated SNPs are in linkage equilibrium. Although the effective number of independent markers in standard GWAS chips has sometimes been assumed to as large as 200,000 (e.g. ref[17]), we believe that 60,000 is a more appropriate value, as analyses of LD[29], genomic inflation factors[69] and eigenvalues from principal components analysis[70] have consistently produced estimates close to 60,000 in European populations. Predictions from theory, based upon random mating populations of a given effective size and for given genome length, also come to this number[36]. Thus the appropriate value for *M* is approximately 60,000.

Independent sample In the context of risk prediction an independent sample means a sample from the same population but excluding individuals that are closely related. Necessarily, the individuals in different samples from the same population will share common ancestors, and indeed this distant sharing underpins the efficacy of a risk predictor.

Cross-validation To test the validity of a prediction in the absence of an independent external validation sample, the sample is divided 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.

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.

Cryptic relatedness Cryptic relatedness is when a sample is thought to comprise unrelated individuals based on record pedigree relationships but in fact includes close relatives, for example 2nd cousin or closer.

Conventionally unrelated Individuals from that are not closely related, for example more distantly related than 3rd cousins

Box 1. Quantifying phenotypic variation explained by SNPs

Quantitative traits

The proportion of phenotypic variance explained (R^2) by a predictor of a quantitative trait formed using estimated effects of all markers depends on the number (M) of independent measured genomic variants (e.g., 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, 36, 38]. If all marker effects are assumed to come from the same normal distribution, then approximately

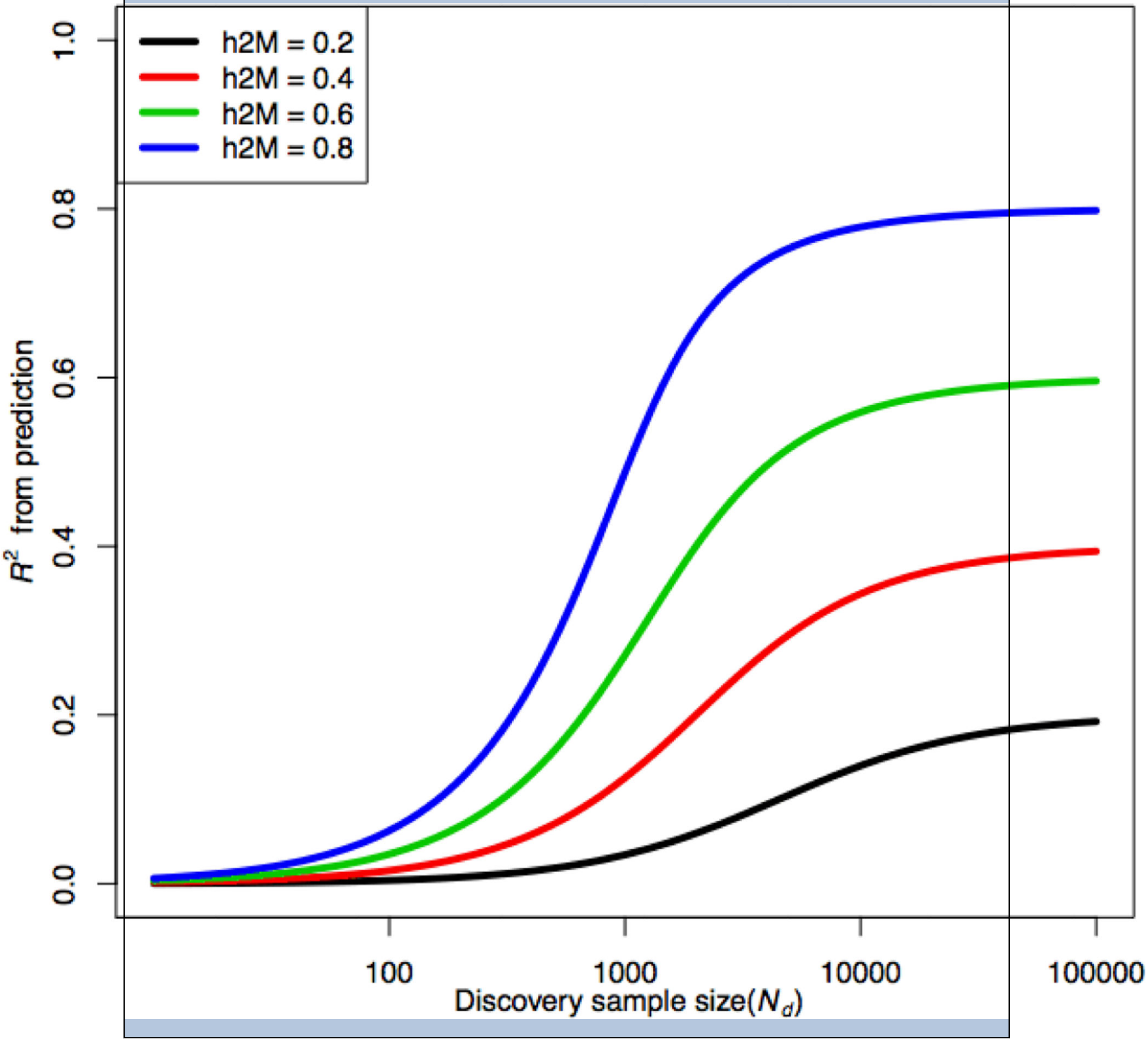
$$R^2 = \frac{h^2_M}{1 + \frac{M}{N_d h^2_M}(1 - R^2)} \qquad \text{[Equation 1]}$$

Equation 1 holds regardless of the genetic architecture of the trait, but we note that the (least squares) 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, estimated, for example, using GCTA[52]. Equation 1 is from the supplement of ref[38]; when R^2 is small it can be ignored from the denominator, otherwise the quadratic in R^2 can be solved. The graph below 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[61].

In this article we use R^2 at 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 (based on 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). Since $R^2_{G,\hat{G}} = \frac{R^2}{h^2}$, 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. Application is usually in case-control validation samples, where the proportion of cases is much higher than in the population. Alternatively, the area under the receiver operator curve (AUC) is reported[74–76], a statistic with a long tradition of use in determining the efficacy of clinical predictors. AUC has a 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 it can be related to other statistics such as R^2_N and AUC[11], has been proposed[77]. Like any estimate on the liability scale, calculation of R_l^2 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.



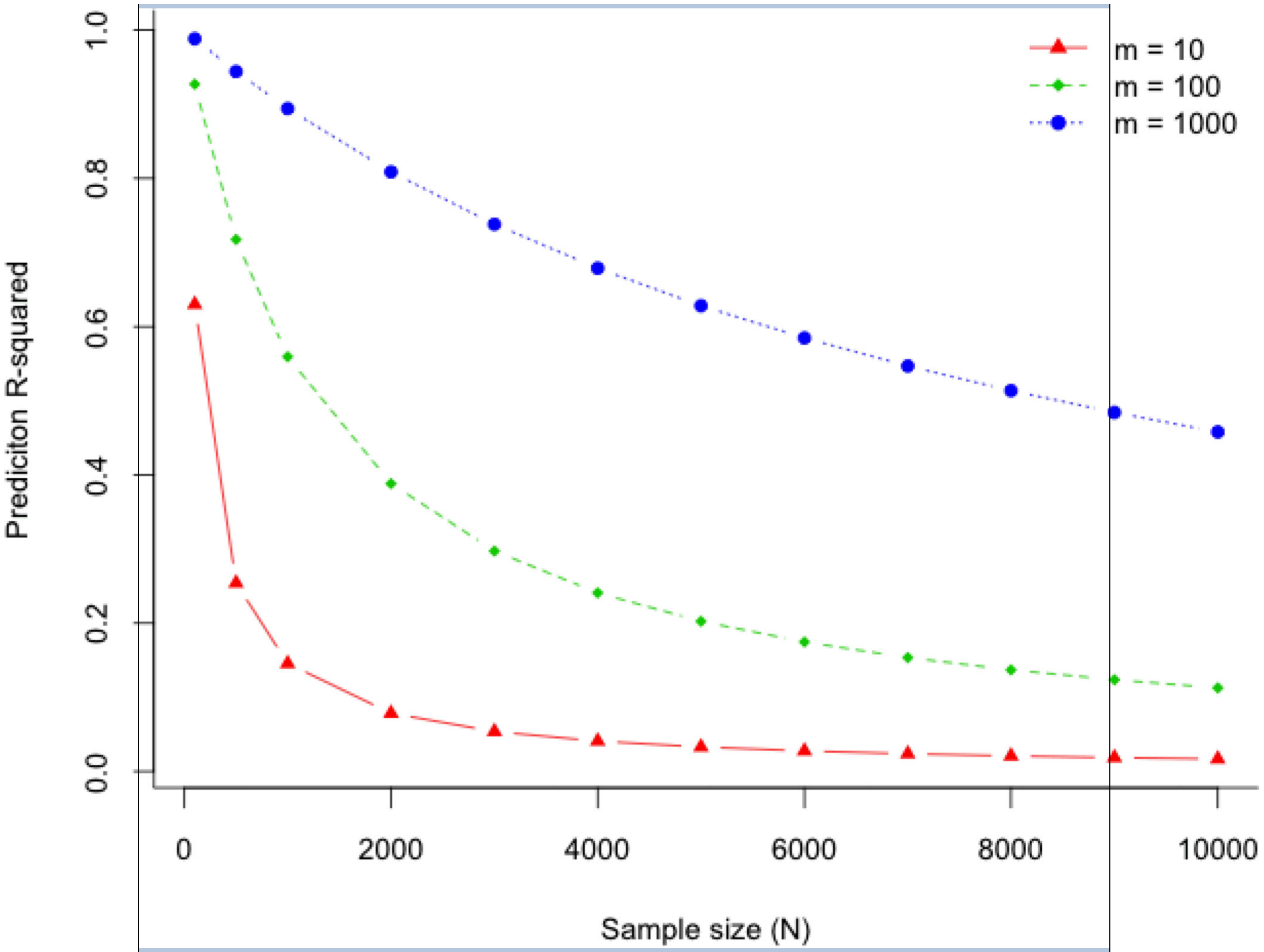
Box 2. Quantifying prediction accuracy for pitfall 2

When discovery and validation samples are independent

When *m* SNPs have been selected from a discovery sample, a simple linear predictor in the validation sample is $\hat{y}=\sum_{i=1}^{m}\hat{b}_i x_i$, with x_i = 0,1 or 2 reference alleles of a SNP and $\hat{b_i}$ 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 squares and more complex Bayesian estimation methods that have been described elsewhere[1, 41, 42]. 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 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 true value of R^2 in the population. The standard error (SE) of *R* is approximately $1/\sqrt{N_v}$ if ρ is very small, and more generally $(1-\rho)^2/\sqrt{N_v}$. In terms of R^2, its SE is approximately $\sqrt{2/N_v}$ with ρ small. A general and a more complicated exact equation was given by Wishart (1931)[77]. 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 the supplementary material 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*, dependent on *m* and assuming *M* = 100,000 independent genomic variants associated with the phenotype is plotted below in which *m* SNPs (*m* = 10, 100 or 1000) 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.



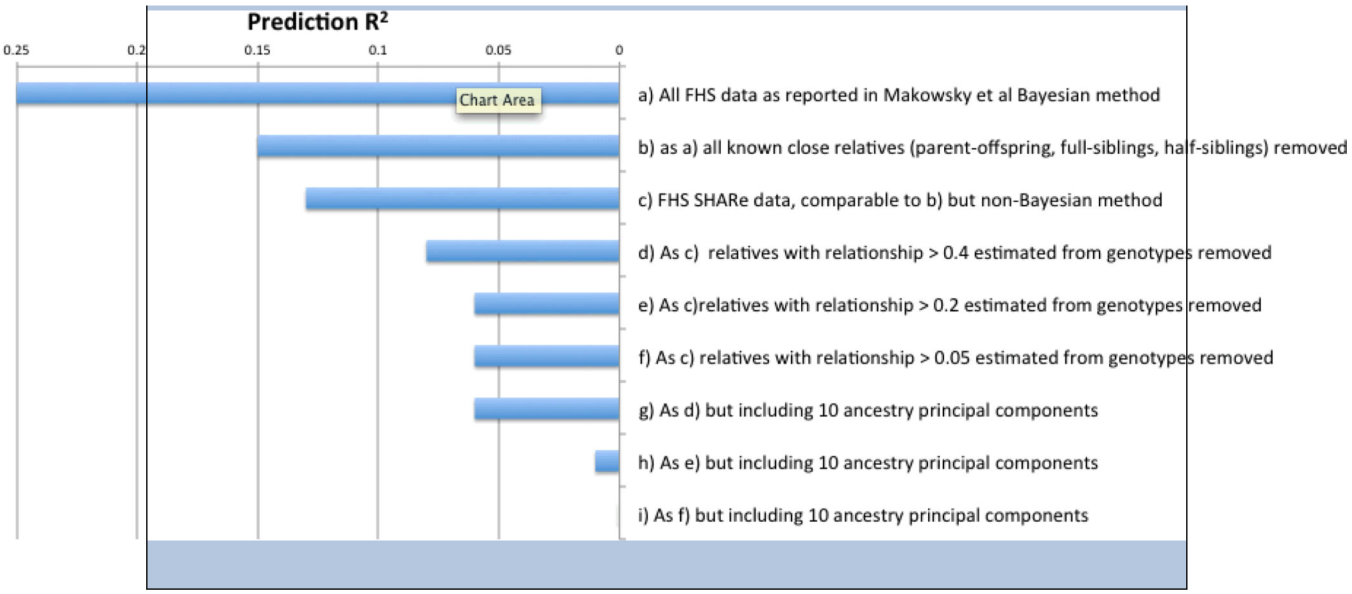
When validation sample overlaps with the discovery sample

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 the supplementary material (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.

Box 3. Using the Framingham Heart Study (FHS) to illustrate pitfalls of validation

The 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[60] showed that including known related individuals in the analysis inflated R^2 (from 0.15 to 0.25) To investigate if 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[61]. We obtained a prediction R^2 of 0.13 using 10-fold cross-validation when restricting to individuals with no known close relatives in the data set based on known pedigree information. (We fit markers individually whereas in the original study[60] markers were fitted simultaneously via a Bayesian random effects model, thus it is expected that a slightly higher R^2 of 0.15 was reported). We repeated the analysis restricting to individuals with pairwise relatedness estimated from the 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[36] or from empirical data on much larger sample sizes[61]. When repeating the analysis using a height phenotype that was adjusted for 10 eigenvectors[66] 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 than the prediction R^2 obtained using unadjusted height. The bulk of the reduction came from correcting for the top eigenvector, representing northwest European vs. southeast European ancestry[63], which is strongly correlated to height (R^2=0.05 in FHS data, consistent with other studies[78, 79]). 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 summarised in the graph below.


Prediction R^2
0.25
0.2
0.15
0.1
0.05
0
Chart Area
a) All FHS data as reported in Makowsky et al Bayesian method
b) as a) all known close relatives (parent-offspring, full-siblings, half-siblings) removed
c) FHS SHARe data, comparable to b) but non-Bayesian method
d) As c) relatives with relationship > 0.4 estimated from genotypes removed
e) As c)relatives with relationship > 0.2 estimated from genotypes removed
f) As c) relatives with relationship > 0.05 estimated from genotypes removed
g) As d) but including 10 ancestry principal components
h) As e) but including 10 ancestry principal components
i) As f) but including 10 ancestry principal components

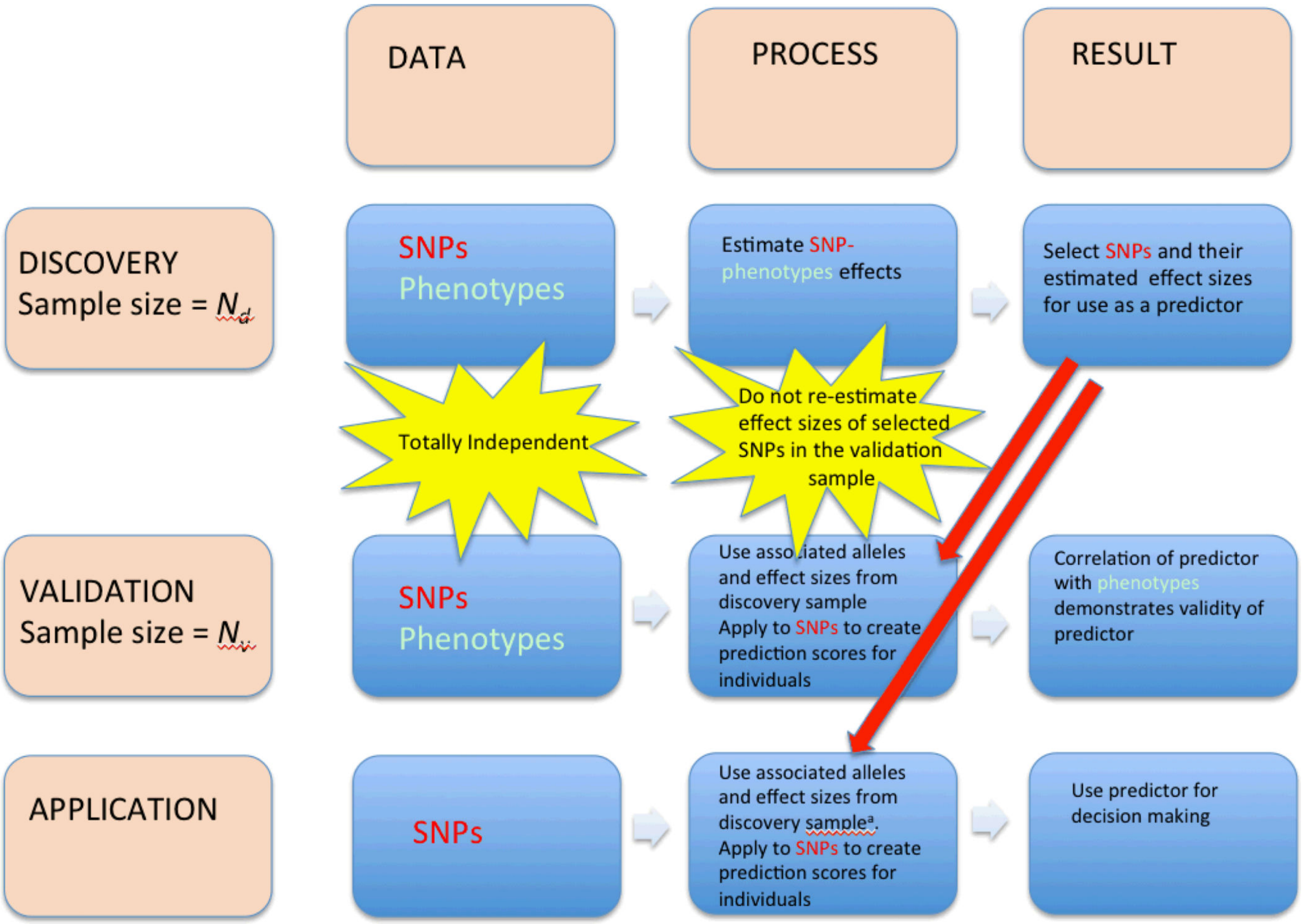
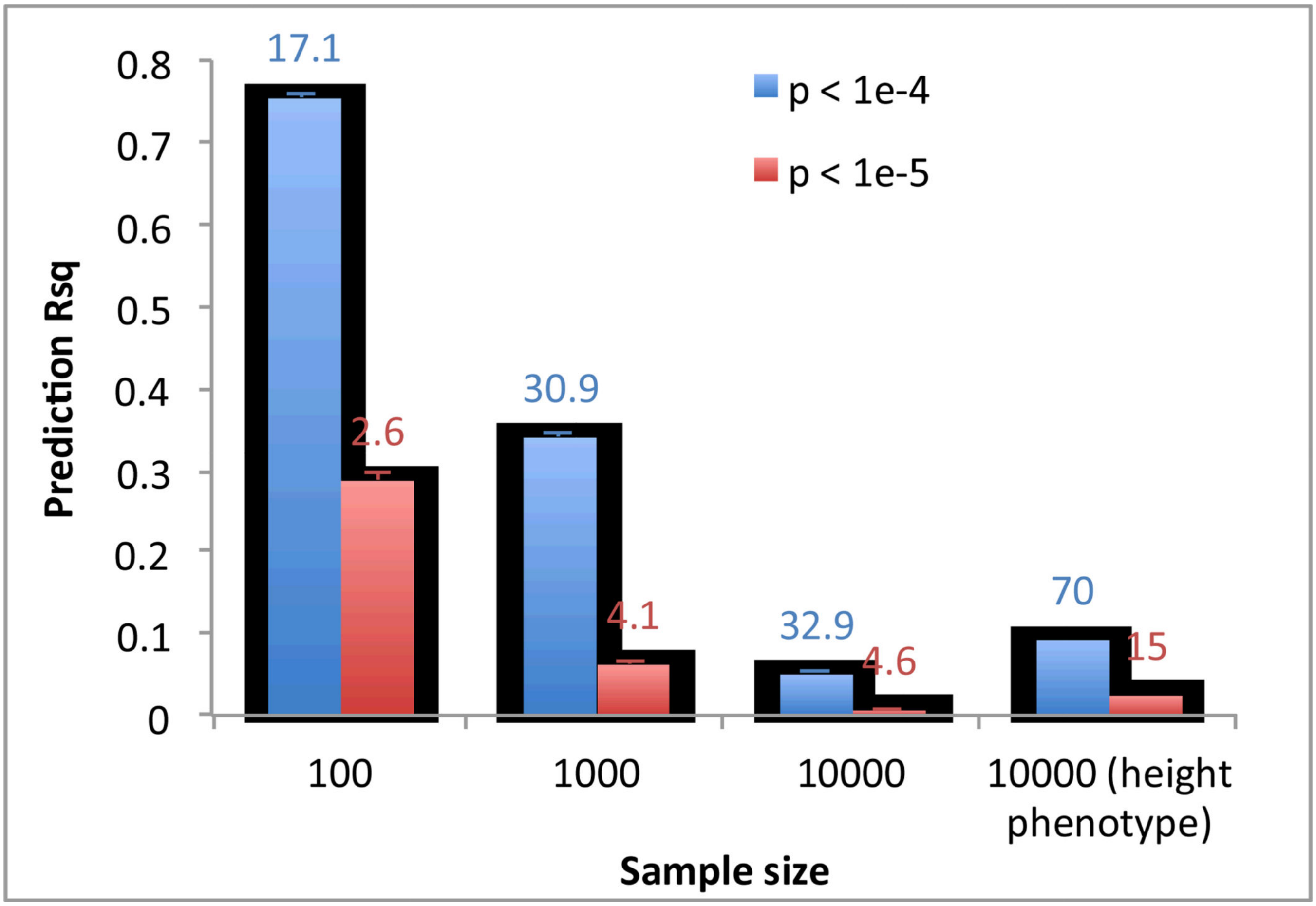
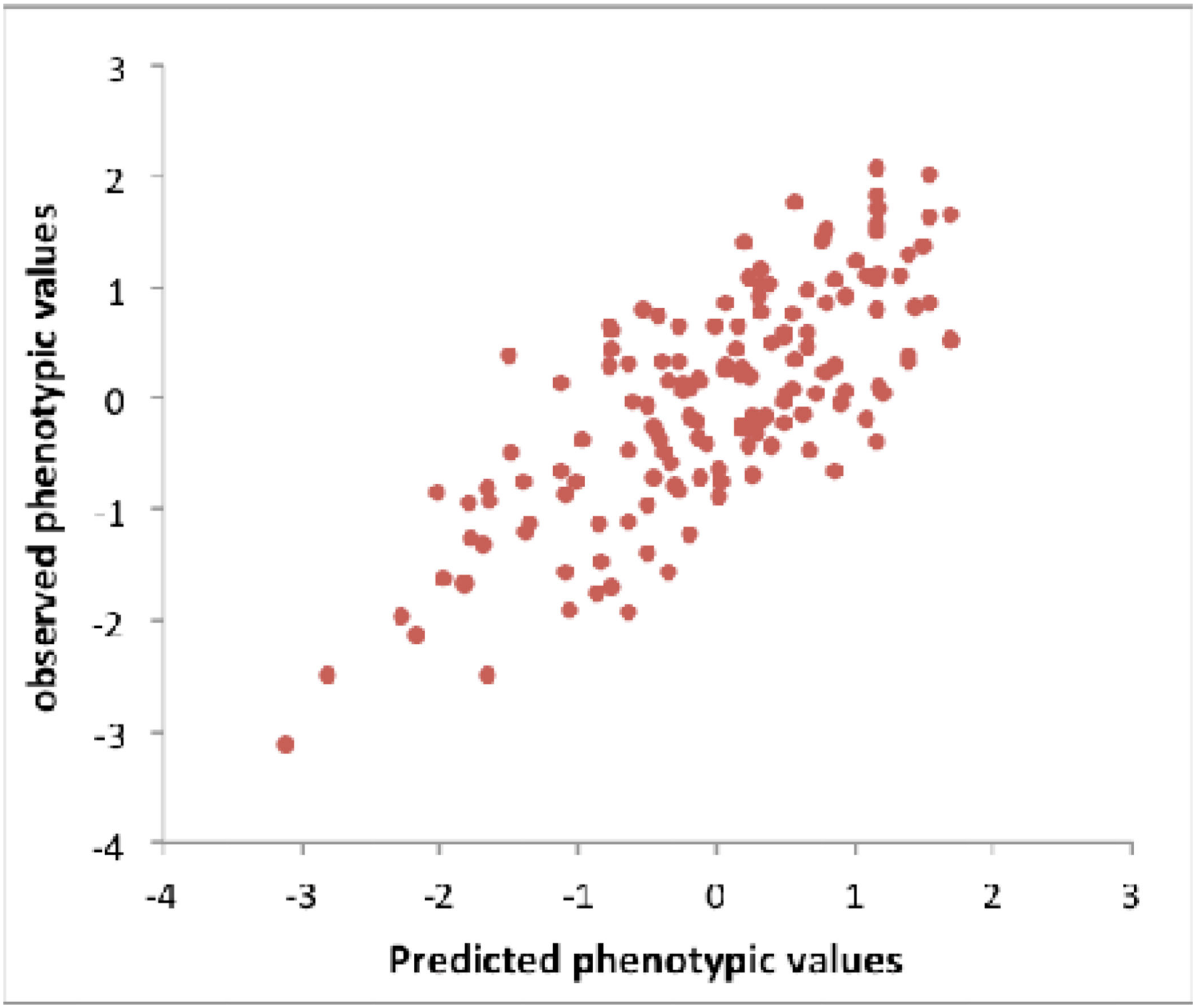

DATA
PROCESS
RESULT
DISCOVERY
Sample size = N_d
SNPs
Phenotypes
Estimate SNP-phenotypes effects
Select SNPs and their estimated effect sizes for use as a predictor
Totally Independent
Do not re-estimate effect sizes of selected SNPs in the validation sample
VALIDATION
Sample size = N_v
SNPs
Phenotypes
Use associated alleles and effect sizes from discovery sample
Apply to SNPs to create prediction scores for individuals
Correlation of predictor with phenotypes demonstrates validity of predictor
APPLICATION
SNPs
Use associated alleles and effect sizes from discovery sample[a].
Apply to SNPs to create prediction scores for individuals
Use predictor for decision making

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 is needed as an input, a process is applied to the data and a result is generated.

[a]. At this stage effect sizes estimated from combined discovery and validation samples can be used.


Prediction Rsq
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
p < 1e-4
p < 1e-5
17.1
2.6
30.9
4.1
32.9
4.6
70
15
100
1000
10000
10000 (height phenotype)
Sample size


observed phenotypic values
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
Predicted phenotypic values

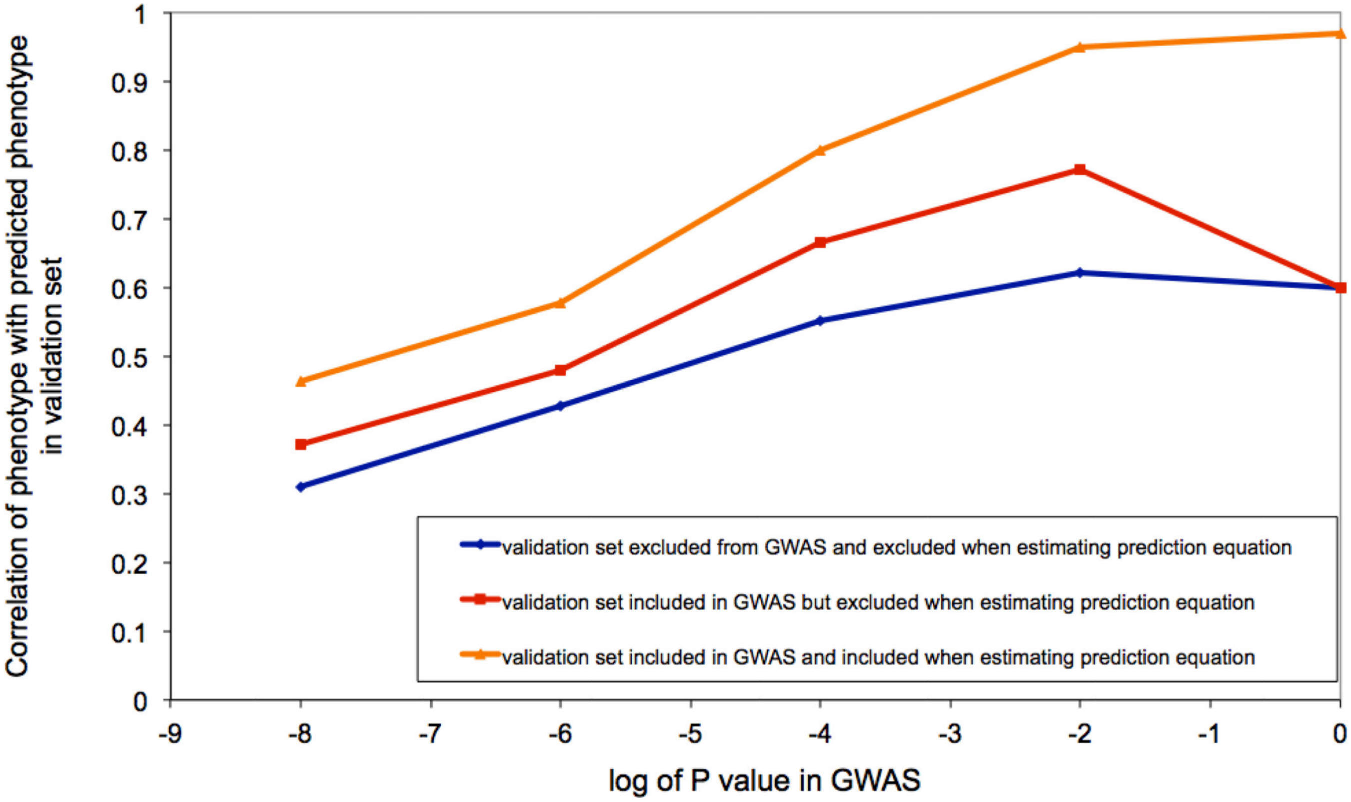


Figure 2. Examples of the overlap pitfall: non-independence of discovery and validation samples

a) Human: High R^2 can be achieved by chance particularly when sample size is small. We simulated GWAS data based upon real human genotype data under the null hypothesis of no association. We used data of 11,586 unrelated European Americans genotyped on 563,212 SNPs [71–73]. 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 performed association analysis for real height phenotype in 10,000 individuals and selected top SNPs at $p < 10^{-5}$ (green bar) and $p < 10^{-4}$ (purple bar) to predict height phenotype in the same sample. The graph shows the mean prediction R^2 over 100 simulation replicates. Error bar: standard error of the mean. The number on top of each column is the mean number of selected SNPs over 100 simulation replicates.

b) *Drosophila*: An example, illustrating bias when selecting the top SNPs. We downloaded genotype data of the *Drosophila* Genetics Reference Panel and simulated phenotypes under the null hypothesis, i.e., random association between each of the > 1 million SNPs and phenotype. We repeated the GWAS analysis reported in[54], selecting the top 10 independently associated SNPs and predicted the phenotypes of the lines using these 10 SNPs. Since in the simulated data there are only random associations between SNP and phenotype any prediction power is false and result of over-fitting. By chance, the top SNPs (in terms of test statistic) explain 57% (R^2=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 (mean of zero and standard deviation of one).

c) Dairy Cattle: The impact of leaving the validation cohort in the discovery set, either at both SNP selection (GWAS) and SNP effect estimation stages, or at the effect size estimation stage only. Data shown are from 2,732 bulls with ~500K 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.