Original Paper

Human Heredity

Hum Hered 2015;80:178–186
DOI: 10.1159/000446581

Published online: September 1, 2016

Accuracy of Gene Scores when Pruning Markers by Linkage Disequilibrium

Frank Dudbridge[a–c] Paul J. Newcombe[b]

[a]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, and [b]MRC Biostatistics Unit, and [c]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

Key Words
Genomic prediction · Polygenic risk score · Genome-wide association study

Abstract
Objective: Gene scores are often used to model the combined effects of genetic variants. When variants are in linkage disequilibrium, it is common to prune all variants except the most strongly associated. This avoids duplicating information but discards information when variants have independent effects. However, joint modelling of correlated variants increases the sampling error in the gene score. In recent applications, joint modelling has offered only small improvements in accuracy over pruning. We aimed to quantify the relationship between pruning and joint modelling in relation to sample size. ***Methods:*** We derived the coefficient of determination R^2 for a gene score constructed from pruned markers, and for one constructed from correlated markers with jointly estimated effects. ***Results:*** Pruned scores tend to have slightly lower R^2 than jointly modelled scores, but the differences are small at sample sizes up to 100,000. If the proportion of correlated variants is high, joint modelling can obtain modest improvements asymptotically. ***Conclusions:*** The small gains observed to date from joint modelling can be explained by sample size. As studies become larger, joint modelling will be useful for traits affected by many correlated variants, but the improvements may remain small. Pruning remains a useful heuristic for current studies.

© 2016 The Author(s)
Published by S. Karger AG, Basel

Introduction

The discovery of thousands of genetic markers for complex traits and the revelation that many more remain to be found have created an emerging discipline of polygenic epidemiology in which the genetic basis of a trait is treated as a single entity [1]. A commonly used, simple yet effective model for this genetic basis is an additive score constructed for each individual as the weighted sum of trait-increasing alleles, with the weights estimated from training data to reflect the relative effect of each marker on the trait [2]. Here, we call this sum a 'gene

KARGER 125

E-Mail karger@karger.com
www.karger.com/hhe

© 2016 The Author(s)
Published by S. Karger AG, Basel
0001–5652/16/0804–0177$39.50/0

Karger Open access

This article is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND) (http://www.karger.com/Services/OpenAccessLicense). Usage and distribution for commercial purposes as well as any distribution of modified material requires written permission.

Frank Dudbridge
Department of Non-Communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine
Keppel Street, London WC1E 7HT (UK)
E-Mail frank.dudbridge@lshtm.ac.uk

score', though various other terms are in use including allele score, genetic risk score, polygenic risk score, or genomic profile score: those terms are perhaps more specific to applications. A gene score may be constructed from a few consistently associated markers, from all nominally significant markers, or from nearly all genotyped markers. For example, to demonstrate that a complex trait has a polygenic basis [3], or to estimate the trait variation explained by a polygenic score [4], it is often optimal to include most of the markers in the score [5]. For Mendelian randomisation, in which it is important that all markers act through the same pathway, it may be preferable to limit the score to individually associated markers [6]. In all cases however, the usual approach is to combine alleles in an additive model that does not allow for dominance within loci or statistical interaction between loci.

Associated markers are often in linkage disequilibrium (LD) with numerous other markers, all of which may show a nominal association. If one marker can account for all the association within a region, perhaps because it is the sole causal variant, then it makes sense to include only that marker in a gene score. On the other hand, if there are several causal variants in mutual LD, then it may be preferable to include all those variants – but only those – in the score. Identifying the casual variants within a region of LD is an important problem when describing aetiology, and there is an extensive literature of statistical methods for this purpose [7–10]. For epidemiological applications such as risk prediction or patient stratification, however, the aim is often to derive a parsimonious but accurate model that is not necessarily aetiological. In this case, a common practice is to discard (*prune*) from the score all those markers whose LD with the most strongly associated marker is above some threshold, perhaps after some initial model selection. A popular algorithm, implemented in the PLINK software, is as follows: identify the most strongly associated marker from univariate analysis; remove all markers whose LD (measured by the squared correlation of coded genotypes, r^2) with this marker is above a threshold; among the remaining markers, identify the most strongly associated marker, and remove all markers in LD with it; repeat until no markers remain. The result of this 'clumping' procedure is a gene score constructed from a considerably reduced number of markers. Since a single marker is used to capture the effect of each LD 'clump', univariate weights can be used, which are often freely available as summary statistics from consortium studies.

An additive score with pruned markers is easily criticised. Some accuracy must be lost by omitting dominance or interaction terms that might exist in truth or by discarding markers with independent effects. Therefore, some efforts have been made to avoid pruning by accounting for LD, when estimating weights for all markers simultaneously [11], or by combining the marker selection with the weight estimation steps [12, 13]. Surprisingly, however, these more advanced methods often achieve no more than a small improvement in accuracy over the basic gene score. The cases in which advanced methods do achieve substantial improvement tend to be in HLA-associated diseases, in which there are multiple strong associations in high LD. Such examples arguably depart from the classical polygenic model of infinitesimal effects, and it is in cases closer to that model that the basic gene score, despite its obvious shortcomings, tends to perform surprisingly well.

One reason that a pruned gene score could have comparable accuracy to an unpruned score is that when training the score from finite data, a score containing fewer markers will have less sampling error than one containing more markers, simply because there are fewer parameters to estimate. Furthermore, the univariate effects in a pruned score are in fact marginal effects that include, to some extent, the effects of correlated markers that have been discarded. The loss of information from pruning depends on the degree of LD with the discarded markers and the effect sizes of those markers, but is offset by the reduced sampling error in the pruned score. It is, therefore, not a given that, with finite training data, an unpruned score with adjustment for LD will have greater accuracy than a pruned score.

Here, we give an analytical description of the effect of pruning, under a simple model in which a proportion of markers with effects on a trait are each in LD with a secondary marker with an independent effect on the trait. We compare a pruning approach, in which only the more strongly associated marker is retained with its marginal weight, to an unpruned approach, in which both markers are retained with their weights as their conditional effects. Our analysis sheds light on the competitive performance of the pruned gene score in polygenic traits and shows that criticisms of additive models as inadequate are not quite accurate, because marginal effects include some of the non-additive signal. By considering the sample size of the training data, we offer new interpretations of recent methods accounting for LD, and we suggest the order of sample size under which advanced methods can be expected to provide more substantial gains.

Methods

We consider the accuracy of a gene score S in terms of its coefficient of determination on a trait Y. When no other predictors of Y are modelled, the coefficient of determination is simply the squared correlation

$$R^2 = \frac{\mathrm{cov}\left(S, Y\right)^2}{\mathrm{var}\left(S\right)\mathrm{var}\left(Y\right)}.$$

Most other measures of accuracy can be expressed in terms of R^2, and it can be readily converted into several alternative quantities when Y is binary [14].

Let the gene score be a sum of contributions from a large number m of independent genomic regions, with each region containing a number of correlated markers

$$S = \sum_{i=1}^{m}\sum_{j}\beta_{ij}X_{ij},$$

where X_{ij} is a numerical code for the genotype of marker j in region i, and β_{ij} is a scalar effect. We assume that the β_{ij}s are independent and identically distributed, and for simplicity (but without loss of generality) that the Ys and X_{ij}s are all standardised. We consider the linear model

$$E(Y \mid S) = S.$$

Consider a region in which there is one marker with an effect on Y. Suppressing index i, the covariance between Y and the fitted values from the single marker model is

$$\mathrm{cov}(\beta_1 X_1, Y) = \mathrm{cov}(\beta_1 X_1, \beta_1 X_1) = \mathrm{var}(\beta_1 X_1) = \beta_1^2. \quad (1)$$

Now suppose there is a second marker in the region, X_2, which also has an effect on Y

$$E(Y \mid X_1, X_2) = \beta_1 X_1 + \beta_2 X_2.$$

The marginal effect of X_1 is now a function of the two SNP effects

$$\beta_1^m = \mathrm{cov}(X_1, Y) = \beta_1 + r\beta_2,$$

where r is the (signed) correlation between X_1 and X_2. The covariance between Y and the fitted values from the marginal single marker model is now

$$\mathrm{cov}(\beta_1^m X_1, Y) = \mathrm{var}(\beta_1^m X_1) = \beta_1^2 + 2r\beta_1\beta_2 + r^2\beta_2^2. \quad (2)$$

Now, we construct a gene score S_{marg} by including only one effect from each of the m regions. Suppose a proportion of regions, π_2, contains two markers with independent effects on Y, a proportion, π_1, contains one marker affecting Y, and the remainder, π_0, contains no markers affecting Y. In regions containing two independent effects, marker X_1 is chosen when $\mid \beta_1^m \mid \geq \mid \beta_2^m \mid$, with X_2 chosen otherwise. In regions containing one effect, the corresponding marker is chosen, and in regions containing no effects, a random marker is chosen with $\beta_1^m = 0$. (In practice, the marker selection will depend on estimated effects; for convenience, we ignore sampling variation and selection bias here.) Consider β_1 and β_2 as random effects, drawn independently from a common distribution F with variance σ^2. Using equations 1 and 2, the covariance between Y and the gene score S_{marg} composed of marginal effects is

$$\mathrm{cov}\left(S_{marg}, Y\right) = m\left[2\pi_2\int_{-\infty}^{\infty}\int_{-\beta_1}^{\beta_1}\left[\beta_1^2 + 2r\beta_1\beta_2 + r^2\beta_2^2\right]dF\left(\beta_2\right)dF\left(\beta_1\right) + \pi_1\sigma^2\right],$$

the factor of 2 in the first term reflecting symmetry in contributions from $\mid \beta_1^m \mid \geq \mid \beta_2^m \mid$ and $\mid \beta_2^m \mid > \mid \beta_1^m \mid$. Since β_1 and β_2 are assumed independent,

$$\mathrm{cov}\left(S_{marg}, Y\right) = m\left[8\pi_2\int_0^{\infty}\int_0^{\beta_1}\left[\beta_1^2 + r^2\beta_2^2\right]dF\left(\beta_2\right)dF\left(\beta_1\right) + \pi_1\sigma^2\right]. \quad (3)$$

Suppose the effects β have been estimated from a sample of size n. Then, the estimated gene score has variance

$$\mathrm{var}(\hat{S}_{marg}) = \mathrm{var}(S_{marg} + \varepsilon_{marg}) = \mathrm{var}(S_{marg}) + \mathrm{var}(\varepsilon_{marg}) = \mathrm{cov}(S_{marg}, Y) + \mathrm{var}(\varepsilon_{marg}),$$

where $\mathrm{var}(\varepsilon_{marg})$ denotes the sampling variance of the estimated gene score. Since the gene score includes one marker per independent region, the sampling variance is equal to the residual variance divided by n, and we have

$$\begin{aligned}&\mathrm{var}\left(\hat{S}_{marg}\right) = \mathrm{cov}\left(S_{marg}, Y\right) + \\ &mn^{-1}\left[\pi_2\left(1 - 8\int_0^{\infty}\int_0^{\beta_1}\left[\beta_1^2 + r^2\beta_2^2\right]dF\left(\beta_2\right)dF\left(\beta_1\right)\right) + \right. \\ &\left.\pi_1\left(1-\sigma^2\right) + \pi_0\right] \\ &= \mathrm{cov}\left(S_{marg}, Y\right) + \\ &mn^{-1}\left[1 - 8\pi_2\int_0^{\infty}\int_0^{\beta_1}\left[\beta_1^2 + r^2\beta_2^2\right]dF\left(\beta_2\right)dF\left(\beta_1\right) - \pi_1\sigma^2\right] \\ &= \mathrm{cov}\left(S_{marg}, Y\right)\left(1 - mn^{-1}\right) + mn^{-1}.\end{aligned} \quad (4)$$

Using equations 3 and 4, the coefficient of determination for the marginal model is

$$R^2_{marg} = \frac{\mathrm{cov}\left(\hat{S}_{marg}, Y\right)^2}{\mathrm{var}\left(\hat{S}_{marg}\right)} = \frac{\mathrm{cov}\left(S_{marg}, Y\right)^2}{\mathrm{cov}\left(S_{marg}, Y\right)\left(1 - mn^{-1}\right) + mn^{-1}}.$$

For comparison, we also consider the case in which one random marker is chosen from each correlated pair. Denoting this gene score by S_{rand}, along the same lines as above, we have

$$\begin{aligned}&\mathrm{cov}\left(S_{rand}, Y\right) = m\left(\pi_2\sigma^2\left(1 + r^2\right) + \pi_1\sigma^2\right) \\ &\mathrm{var}\left(\hat{S}_{rand}\right) = \mathrm{cov}\left(S_{rand}, Y\right) + mn^{-1}\left(\pi_2\left(1 - \sigma^2\left(1 + r^2\right)\right) + \right. \\ &\left.\pi_1\left(1 - \sigma^2\right) + \pi_0\right) \\ &R^2_{rand} = \frac{\mathrm{cov}\left(S_{rand}, Y\right)^2}{\mathrm{var}\left(\hat{S}_{rand}\right)}.\end{aligned}$$

Now, consider a gene score S_{joint}, in which both effects are included from regions with two independent effects, with other regions treated identically to S_{marg}. For a region with two effects, the covariance between Y and the fitted values from a joint model is

$$\mathrm{cov}(\beta_1 X_1 + \beta_2 X_2, Y) = \mathrm{var}(\beta_1 X_1 + \beta_2 X_2) = \beta_1^2 + 2r\beta_1\beta_2 + \beta_2^2. \quad (5)$$

Recalling that β_1 and β_2 are assumed independent, the covariance between Y and S_{joint} from the joint model easily follows as

$$\mathrm{cov}(S_{joint}, Y) = \mathrm{var}(S_{joint}) = m(2\pi_2 + \pi_1)\sigma^2.$$

Standard linear regression theory gives the variance-covariance matrix of $(\hat{\beta}_1, \hat{\beta}_2)'$ as the residual variance times $(\mathbf{X}'\mathbf{X})^{-1}$,

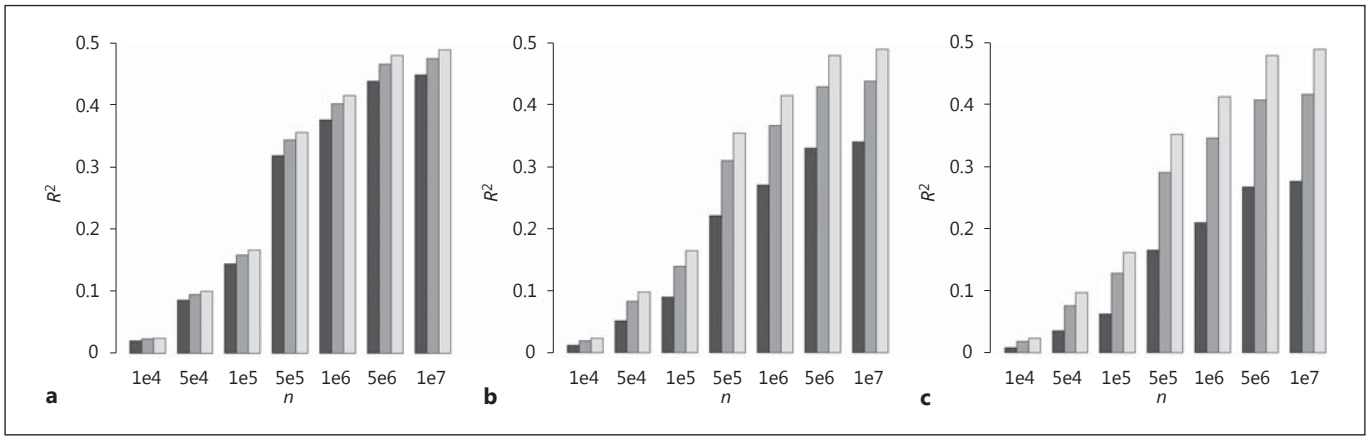

Fig. 1. R^2 for the prediction of trait Y using gene scores estimated from samples of size n. Of 10^5 independent genomic regions, a proportion π_2 contains two markers affecting Y, with correlation $r^2 = 0.1$ between each pair, a proportion π_1 contains one marker affecting Y, and a proportion $\pi_0 = 0.95$ contains no markers affecting Y. **a** π_2:π_1 = 1:9; **b** π_2:π_1 = 1:1; **c** π_2:π_1 = 9:1. Dark bars = One random marker per correlated pair R^2_{rand}. Grey bars = Marker with the strongest absolute effect per pair R^2_{marg}. Light bars = Both markers per pair R^2_{joint}.

where $\mathbf{X}$ is the design matrix. Using equation 5, and recalling that all variables are standardised, we write this as

$$\left(1-\beta_1^2-2r\beta_1\beta_2-\beta_2^2\right)n^{-1}\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}^{-1}$$

$$=\left(1-\beta_1^2-2r\beta_1\beta_2-\beta_2^2\right)n^{-1}\left(1-r^2\right)^{-1}\begin{pmatrix}1 & -r\\ -r & 1\end{pmatrix}$$

from which the sampling variance of $\hat{\beta}_1X_1 + \hat{\beta}_2X_2$ is

$2\,(1 - \beta_1^2 - 2r\beta_1\beta_2 - \beta_2^2)\,n^{-1}$.

Summing this sampling variance over the $m\pi_2$ regions with two effects, we obtain the variance of the estimated gene score as

$$\mathrm{var}(\hat{S}_{joint}) = \mathrm{cov}(S_{joint}, Y) + mn^{-1}\,[2\pi_2\,(1 - 2\sigma^2) + \pi_1\,(1 - \sigma^2) + \pi_0].$$

The coefficient of determination for the joint model is

$$R^2_{joint} = \frac{m\left(2\pi_2+\pi_1\right)^2\sigma^4}{\left(2\pi_2+\pi_1\right)\sigma^2+n^{-1}\left[2\pi_2\left(1-2\sigma^2\right)+\pi_1\left(1-\sigma^2\right)+\pi_0\right]}.$$

From the above it is clear that the relative accuracy of pruned gene scores, constructed from marginal models, and unpruned gene scores, constructed from joint models, depends upon several parameters: the sample size n, total number of regions m, proportions of regions containing two (π_2) and one (π_1) marker in the joint models, squared correlation r^2 between genotypes in regions containing two markers, and distribution F of genetic effects.

We explored the accuracy of pruned and unpruned gene scores under parameters typical of current studies. The sample size was varied between 10^4 and 10^6. The number of regions was fixed at 10^5, of which $\pi_0 = 0.95$ contained no marker affecting the trait [4, 15]. Of the remainder, we varied the odds π_2:π_1 to reflect different proportions of regions with one or two independent effects. For the correlation, we considered $r^2 = 0.1$ and $r^2 = 0.2$, values commonly used when pruning markers [3, 4, 15], and $r^2 = 0.95$, a value sometimes used to prune markers before a joint analysis in order to reduce collinearity [9]. Finally, for the distribution of genetic effects, we followed several authors in assuming a normal distribution on the standardised genotype scale [5, 9, 11, 16, 17] with mean zero and total genetic variance 0.5. This was distributed equally across all markers with effects, so

$$m\,(2\pi_2 + \pi_1)\,\sigma^2 = 0.5$$
$$\sigma^2 = (2m\,(2\pi_2 + \pi_1))^{-1}.$$

In addition, we considered scenarios with more than two markers with effects in regions of LD. We assumed again that the proportion of regions $\pi_0 = 0.95$ had no markers with effects and $\pi_1 = 0.1 \times (1 - \pi_0)$ had one marker with an effect. Among the remaining proportion $\pi_2 = 0.9 \times (1 - \pi_0)$, we considered scenarios where all those regions had k markers with effects, for $k = 3, 4, 5$. We assumed that $r^2 = 0.1$ between each pair of markers with effects. We considered just the asymptotic case $n \to \infty$, so we may ignore the sampling variation in the estimated gene score. Under these assumptions, equation 4 for the marginal gene score S_{marg} easily generalises to $k > 2$; we evaluated the integral by Monte Carlo quadrature with 10^6 draws for each variable. For the joint gene score, R^2_{joint} tends to the total genetic variance, here 0.5, for large n. The random effects variance becomes

$$m\,(k\pi_2 + \pi_1)\,\sigma^2 = 0.5$$
$$\sigma^2 = (2m\,(k\pi_2 + \pi_1))^{-1}.$$

Results

For within region correlation $r^2 = 0.1$, figure 1 and table 1 show the coefficients of determination for random, marginal and joint gene scores when 10, 50 and 90% of

Table 1. R^2 for the prediction of trait Y using gene scores estimated from samples of size n

π_2:π_1	n	1×10^4	5×10^4	1×10^5	5×10^5	1×10^6	5×10^6	1×10^7
1:9	R^2_{rand}	0.020	0.085	0.144	0.319	0.377	0.440	0.449
	R^2_{marg}	0.023	0.095	0.159	0.344	0.403	0.467	0.476
	R^2_{joint}	0.024	0.100	0.166	0.357	0.416	0.481	0.490
1:1	R^2_{rand}	0.012	0.051	0.087	0.221	0.271	0.331	0.340
	R^2_{marg}	0.019	0.082	0.139	0.310	0.367	0.429	0.439
	R^2_{joint}	0.023	0.098	0.164	0.355	0.415	0.480	0.490
9:1	R^2_{rand}	0.008	0.035	0.062	0.164	0.210	0.267	0.277
	R^2_{marg}	0.018	0.076	0.128	0.291	0.346	0.408	0.418
	R^2_{joint}	0.023	0.097	0.162	0.353	0.414	0.480	0.490

Of 10^5 independent genomic regions, a proportion π_2 contains two markers affecting Y, with correlation r^2 = 0.1 between each pair, a proportion π_1 contains one marker affecting Y, and a proportion π_0 = 0.95 contains no markers affecting Y.

Table 2. R^2 for the prediction of trait Y using gene scores estimated from samples of size n

π_2:π_1	n	1×10^4	5×10^4	1×10^5	5×10^5	1×10^6	5×10^6	1×10^7
1:9	R^2_{rand}	0.020	0.087	0.146	0.323	0.281	0.444	0.454
	R^2_{marg}	0.023	0.096	0.160	0.346	0.405	0.468	0.478
	R^2_{joint}	0.024	0.100	0.166	0.357	0.416	0.481	0.490
1:1	R^2_{rand}	0.013	0.056	0.097	0.235	0.287	0.347	0.357
	R^2_{marg}	0.020	0.084	0.142	0.316	0.373	0.436	0.445
	R^2_{joint}	0.023	0.098	0.164	0.354	0.415	0.480	0.490
9:1	R^2_{rand}	0.009	0.040	0.071	0.186	0.232	0.291	0.300
	R^2_{marg}	0.018	0.078	0.133	0.299	0.355	0.417	0.426
	R^2_{joint}	0.023	0.097	0.162	0.352	0.414	0.480	0.490

Within regions with two markers affecting Y, the correlation is r^2 = 0.2 between each pair. Other details as in table 1.

the regions with effects harbour two independent effects that may warrant joint modelling. When few regions have two independent effects, the marginal gene score is almost as accurate as the joint gene score, as we might expect; although R^2_{joint} is always greater than R^2_{marg}, the difference is always within 2%. When half of the regions have two independent effects, the differences remain small, <3% up to $n = 10^5$ and just over 5% at $n = 10^7$. More substantial differences occur when most of the regions have two independent effects: in the large sample limit, the difference in R^2 is about 8%, although clear differences only emerge with samples of order $n = 10^5$, when R^2 exceeds 10% in absolute value. Comparing the random to the marginal gene score, there is a clear benefit in selecting the marker with the strongest effect when two effects are present.

For r^2 = 0.2, the results are qualitatively similar (table 2). These results suggest that under the common strategy of retaining the strongest effect in a region and pruning markers with $r^2 < 0.2$, little predictive accuracy is lost unless a high proportion of regions has multiple independent effects. Even then, the loss of information from pruning is small at sample sizes less than about 100,000.

The results for r^2 = 0.95 are shown in figure 2. For all proportions of regions with two independent effects, R^2_{marg} was within 1% of R^2_{joint}, and in fact slightly exceeded it. These results suggest, as might be expected, that there is little loss (in fact, in finite samples a possible gain) of

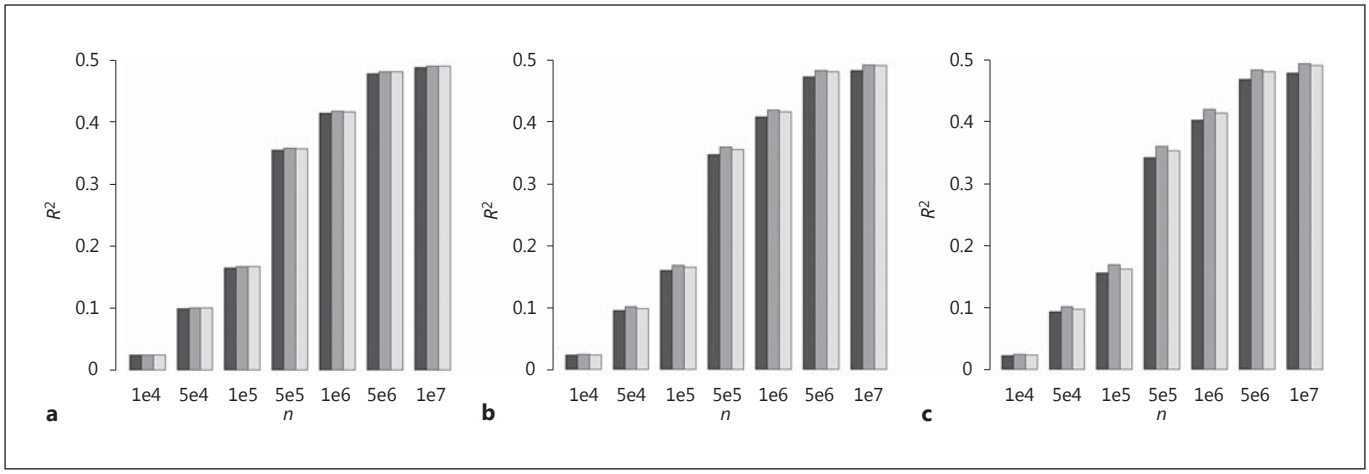

Fig. 2. R^2 for the prediction of trait Y using gene scores estimated from samples of size n. Within regions with two markers affecting Y, the correlation is $r^2 = 0.95$ between each pair. **a** π_2:π_1 = 1:9; **b** π_2:π_1 = 1:1; **c** π_2:π_1 = 9:1. Dark bars = One random marker per correlated pair R^2_{rand}. Grey bars = Marker with the strongest absolute effect per pair R^2_{marg}. Light bars = Both markers per pair R^2_{joint}. Other details as in figure 1.

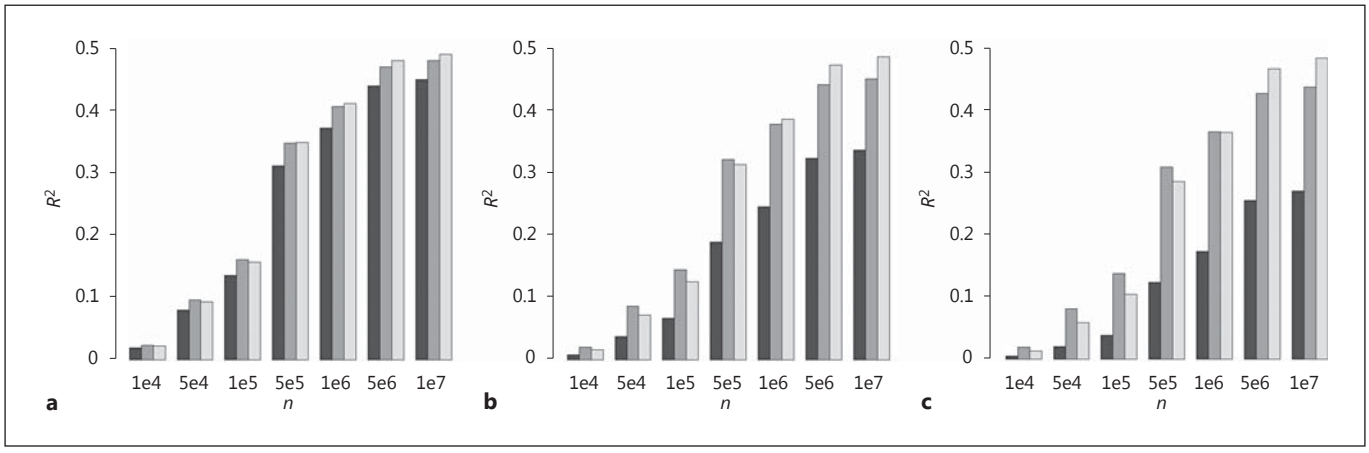

Fig. 3. R^2 for the prediction of trait Y using gene scores estimated from samples of size n. All regions contain markers affecting Y, $\pi_0 = 0$. **a** π_2:π_1 = 1:9; **b** π_2:π_1 = 1:1; **c** π_2:π_1 = 9:1. Dark bars = One random marker per correlated pair R^2_{rand}. Grey bars = Marker with the strongest absolute effect per pair R^2_{marg}. Light bars = Both markers per pair R^2_{joint}. Other details as in figure 1.

information when pruning markers with high r^2 to reduce collinearity in fitting a joint model.

In figure 3 and table 3, we show the results for $r^2 = 0.1$ under the classical polygenic model in which all markers have effects, $\pi_0 = 0$. Compared to $\pi_0 = 0.95$ (fig. 1, table 1), the joint model has less accuracy at all sample sizes, because the effect size for each marker is smaller in relation to the sampling error. Consequently, the marginal model has greater accuracy than the joint model at some sample sizes. Although higher values of π_0 are more consistent with recent data [4, 15, 17], this result shows that the marginal and joint models can have similar accuracy for more highly polygenic traits.

Considering greater numbers of markers with effects in regions of LD, the asymptotic R^2_{marg} was approximately 0.42 for $k = 2$ markers in each such region, 0.37 for $k = 3$, 0.33 for $k = 4$, and 0.30 for $k = 5$.

Table 3. R^2 for the prediction of trait Y using gene scores estimated from samples of size n

π_2:π_1	n	1×10^4	5×10^4	1×10^5	5×10^5	1×10^6	5×10^6	1×10^7
1:9	R^2_{rand}	0.018	0.079	0.135	0.310	0.370	0.438	0.448
	R^2_{marg}	0.023	0.096	0.160	0.346	0.405	0.469	0.479
	R^2_{joint}	0.022	0.093	0.156	0.347	0.410	0.479	0.489
1:1	R^2_{rand}	0.008	0.037	0.066	0.188	0.245	0.322	0.336
	R^2_{marg}	0.020	0.086	0.144	0.320	0.378	0.440	0.450
	R^2_{joint}	0.016	0.071	0.125	0.313	0.385	0.472	0.485
9:1	R^2_{rand}	0.004	0.020	0.038	0.123	0.173	0.253	0.269
	R^2_{marg}	0.019	0.081	0.137	0.307	0.363	0.426	0.435
	R^2_{joint}	0.013	0.058	0.104	0.284	0.362	0.465	0.482

All regions contain markers affecting Y, $\pi_0 = 0$. Other details as in table 1.

Discussion

Pruning markers according to LD is a common practice when constructing gene scores either from a limited number of associated regions or from the whole genome. It is particularly convenient when using summary statistics from consortium studies to obtain the marker weights. The intention is to avoid duplication of information within the score, but the attendant concern is that informative markers may be discarded. Although methods are available to dissect regions of LD into independent signals, their power is lower than the univariate analysis, and it may not be easy to automate their application across genome-wide datasets. Here, we note that the marginal effect of a marker includes some of the effect of pruned markers in LD, and demonstrate the trade-off between the pruning and the sampling error. Although asymptotically it is always preferable to include all markers and adjust for LD, at finite sample size it may be preferable to prune markers, discarding information but reducing the sampling error. As a rule of thumb, under current models of polygenic traits, in which about 5% of all markers have effects, the marginal score has similar accuracy to the joint score for sample sizes up to 100,000, unless a large number of regions have two or more independent effects.

Our results shed light on some recent findings. Estimates of disease liability explained by genome-wide markers are similar between methods that use pruned markers [4, 15] and those that use all markers [18, 19], though the latter tend to be slightly higher. The prediction accuracy of a gene score accounting for LD across the genome is not much higher than that of one based on pruned markers [11]. As the sample sizes in these studies are of order 10^4, these findings are consistent with our results. The exceptions are in diseases with a strong HLA component, such as type 1 diabetes, rheumatoid arthritis, and multiple sclerosis. In these cases, there are likely to be multiple risk loci with relatively strong effects within extended regions of LD. These genetic models depart from the polygenic model considered here, which assumes a normal distribution of effects and independence of effects within regions of LD. There is a stronger case for accounting for LD within the HLA region when such effects exist, but at current sample sizes, standard pruning approaches appear adequate. However, as sample sizes approach order 10^6, it will become more important to allow for LD to fully exploit the available information.

The marginal model selecting the strongest effect in each region performs remarkably well. Indeed, for the situation in which $r^2 = 0.1$ and π_2:π_1 is 9:1, nearly half of the genetic effects are discarded, and the marginal effects of the retained markers are at most $(1+\sqrt{0.1}) = 1.32$ times their conditional effects. Yet the marginal model has asymptotic R^2 at 85% of the joint model (table 1, bottom two rows). Under a liability threshold model, this corresponds to an area under the receiver-operator characteristic curve (AUC) of 85 and 87% for the marginal and joint models, respectively, for a disease with prevalence 10% [5], or of 92 and 94% for a disease with prevalence 1%. This surprising result occurs because the selection of the strongest effect requires inspection of both effects, which then both contribute to the marginal effect. This approach is, thus, implicitly performing a bivariate analysis, whose accuracy is closer to the explicit bivariate analysis than might be intuitively expected.

When there were more markers with effects in each region of LD, the marginal model became less accurate compared to the joint model. However, our model was designed to represent a scenario close to the worst case: 90% of the regions contained multiple effects, and the r^2 between each pair of markers with effects was low at 0.1. In reality, the number of effects in each region would vary, as would the r^2 between pairs of markers with effects. Under the worst case we considered, with 5 markers with effects in 90% of the regions, the asymptotic R^2 of the marginal model was about 60% of that of the joint model, corresponding to an AUC of 80% for a disease with prevalence 10%, or 87% for a disease with prevalence 1%. In practice, therefore, we might not expect very large increases in the AUC when moving from pruned to jointly modelled gene scores.

We note some limitations of our model. The genome does not consist of independent regions, in which, if two markers affect the trait, the correlation between their genotypes is constant. However, comparing table 1 and 2, we see that when increasing r^2 between retained and pruned markers, the accuracy of the pruned scores increases. Therefore, fixing r^2 in each table provides a lower bound on what would be observed if we allowed r^2 to be at least the value fixed. In this sense, our results are conservative. We also assume that the marginal analysis always selects the marker with the greater true effect, and also that the joint analysis always selects the two markers with effects when there are two such markers, and the one marker with an effect when there is just one. In practice, these selections will be affected by sampling variation, and some form of model comparison may be required in the joint model. The R^2 values will be lower for all models, but there may be less difference in performance between them. However, the asymptotic results will be the same, and we expect our qualitative conclusions to be unchanged. We further ignore selection bias ('winner's curse') in the estimated effects of the selected markers, assume that effects are independent within regions, and do not consider selecting markers by a p value threshold.

However, our aim is to demonstrate analytically the information loss by pruning and its relationship to sample size. Our idealised model explains the surprisingly good current performance of pruning, while projecting the gains from adjusting for LD in future larger studies. Simulations based on real genotypes have obtained similar results [11], though for a more limited range of sample sizes and without controlling the number of effects within a region of LD. We show that when the results from joint modelling are observed to be similar to those using pruning, it could be explained by a low average number of effects in each region of LD, by a low sample size, or by a highly polygenic model. Conversely, strong differences in performance between pruned and jointly modelled gene scores are suggestive of many regions of LD containing multiple markers with independent effects.

Many fine-mapping studies have identified independent effects within regions of LD [20, 21], with the proportion of such regions being reported as high as one-third [22]. Given the limited power to detect multiple independent effects, the true proportion may be higher. Our results suggest that pruning remains an effective strategy for current studies and will continue to capture a high proportion of heritability in future studies. However, to fully exploit the data, it will become increasingly important to jointly model the effects of correlated markers as sample sizes approach the millions. Light pruning, say to $r^2 = 0.95$, can alleviate problems of collinearity and reduce the size of the model space with minimal loss of information.

Acknowledgements

This work was funded by the MRC (MR/K006215/1).

References

1 Dudbridge F: Polygenic epidemiology. Genet Epidemiol 2016;40:268–272.
2 Wray NR, Lee SH, Mehta D, et al: Research review: polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry 2014;55:1068–1087.
3 Purcell SM, Wray NR, Stone JL, et al: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.
4 Palla L, Dudbridge F: A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am J Hum Genet 2015;97:250–259.
5 Dudbridge F: Power and predictive accuracy of polygenic risk scores. PLoS Genet 2013;9:e1003348.
6 Burgess S, Thompson SG: Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol 2013;42:1134–1144.
7 Cordell HJ, Clayton DG: A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. Am J Hum Genet 2002;70:124–141.
8 Wallace C, Cutler AJ, Pontikos N, et al: Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. PLoS Genet 2015;11:e1005272.

9 Yang J, Ferreira T, Morris AP, 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–S3.
10 Farh KK, Marson A, Zhu J, et al: Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337–343.
11 Vilhjalmsson BJ, Yang J, Finucane HK, et al: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 2015;97:576–592.
12 Abraham G, Kowalczyk A, Zobel J, et al: Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet Epidemiol 2013;37:184–195.
13 Warren H, Casas JP, Hingorani A, et al: Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. Genet Epidemiol 2014;38:72–83.
14 Lee SH, Goddard ME, Wray NR, et al: A better coefficient of determination for genetic profile analysis. Genet Epidemiol 2012;36:214–224.
15 Stahl EA, Wegmann D, Trynka G, et al: Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 2012;44:483–489.
16 Meuwissen TH, Hayes BJ, Goddard ME: Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001;157:1819–1829.
17 Moser G, Lee SH, Hayes BJ, et al: Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet 2015;11:e1004969.
18 Speed D, Balding DJ: MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 2014;24:1550–1557.
19 Lee SH, Wray NR, Goddard ME, et al: Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.
20 Orr N, Dudbridge F, Dryden N, et al: Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet 2015;24:2966–2984.
21 Beecham AH, Patsopoulos NA, Xifara DK, et al: Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 2013;45:1353–1360.
22 Trynka G, Hunt KA, Bockett NA, et al: Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 2011;43:1193–1201.