Human
Heredity

Original Paper

Hum Hered 2020;85:24–34
DOI: 10.1159/000513303

Received: August 14, 2020
Accepted: November 19, 2020
Published online: February 12, 2021

Challenges of Adjusting Single-Nucleotide Polymorphism Effect Sizes for Linkage Disequilibrium

Valentina Escott-Price[a, b] Karl Michael Schmidt[c]

[a]Division of Psychiatry and Clinical Neurosciences, Cardiff University, Cardiff, UK; [b]Dementia Research Institute, Cardiff University, Cardiff, UK; [c]School of Mathematics, Cardiff University, Cardiff, UK

Keywords
Polygenic risk score · Linkage disequilibrium · Tikhonov regularisation

Abstract
Background: Genome-wide association studies (GWAS) were successful in identifying SNPs showing association with disease, but their individual effect sizes are small and require large sample sizes to achieve statistical significance. Methods of post-GWAS analysis, including gene-based, gene-set and polygenic risk scores, combine the SNP effect sizes in an attempt to boost the power of the analyses. To avoid giving undue weight to SNPs in linkage disequilibrium (LD), the LD needs to be taken into account in these analyses. ***Objectives:*** We review methods that attempt to adjust the effect sizes (β-coefficients) of summary statistics, instead of simple LD pruning. ***Methods:*** We subject LD adjustment approaches to a mathematical analysis, recognising Tikhonov regularisation as a framework for comparison. ***Results:*** Observing the similarity of the processes involved with the more straightforward Tikhonov-regularised ordinary least squares estimate for multivariate regression coefficients, we note that current methods based on a Bayesian model for the effect sizes effectively provide an implicit choice of the regularisation parameter, which is convenient, but at the price of reduced transparency and, especially in smaller LD blocks, a risk of incomplete LD correction. ***Conclusions:*** There is no simple answer to the question which method is best, but where interpretability of the LD adjustment is essential, as in research aiming at identifying the genomic aetiology of disorders, our study suggests that a more direct choice of mild regularisation in the correction of effect sizes may be preferable.

Introduction

Genome-wide association studies (GWAS) have quickly advanced the field of genetics of complex genetic disorders, led to generation of large data sets and gave a push to developing novel data analysis methods to summarise the vast amount of information and make use of it in a comprehensible manner. GWAS were successful in identifying genetic variants (single-nucleotide polymorphisms [SNPs]) that show association with a disease, but their individual effect sizes are small and require

karger@karger.com
www.karger.com/hhe

Karger
OPEN ACCESS

Valentina Escott-Price
Dementia Research Institute, Hadyn Ellis Building
Cardiff University, Maindy Road
Cardiff CF24 4HQ (UK)
EscottPriceV @ cf.ac.uk

large sample sizes to achieve statistical significance. Methods of post-GWAS analysis such as gene-based and gene-set analysis and the use of polygenic risk scores (PRSs) combine the SNP effect sizes in an attempt to boost the power of the analyses. The most commonly used approaches of assessing the significance of a gene or gene set are implemented in the MAGMA data analysis tool [1]. In addition, the polygenic risk score approach provides an individual score per person and is often applied to the whole genome. The PRS for an individual is calculated as a linear combination of the number of risk alleles carried by the individual, weighted by the effect size of the SNPs estimated from an independent sample [2]. This mimics the combination of the SNP genotypes in multivariate (logistic) regression, but as the coefficients are taken from single-SNP associations, the linkage disequilibrium (LD) between SNPs needs to be taken into account in order to avoid giving undue weight to correlated SNP genotypes. One way to achieve this is to avoid LD by selecting independent (LD-pruned) SNPs for inclusion in the risk score and, to make the score disease relevant, prioritising the SNPs by their strength of association to the disease (*p* value), a simple approach now often referred to as pruning and thresholding ("P + T"). Several real data and simulation studies have shown that this strategy is limited in power and falls short of the heritability explained by the SNPs [3, 4]. Strategies were developed to adjust the effect sizes for LD, instead of LD pruning, thus preserving information carried by SNP markers in LD [1, 3, 4]. Several methods that have been proposed and are gaining wide-spread use are based on a Bayesian framework, assuming a prior distribution of putative effect sizes and LD information from a reference panel to obtain posterior joint effect sizes from those observed in individual SNP association [4–7]. In earlier work, starting from ideas similar to MAGMA-PCA [1], we have suggested a more direct adjustment of the SNP effect sizes using the singular value decomposition of the SNP-SNP LD matrix [3].

In this paper we analyse the adjustment of SNP effect sizes performed by the methods mentioned above, considering in particular predicted LD (LDpred) as a model indicative of the more complex variants. Observing the similarity of the process with the simple Tikhonov-regularised ordinary least square estimate for the putative multivariate regression coefficients, we consider the questions to what extent LD correction is actually performed, and how transparent the adjustment and its interpretation are to the user. The latter point may be an important element in the approach to understanding the biological mechanisms of complex genetic disorders, as, for example, PRSs are now used not only to test the statistical significance and prediction accuracy in a sample, but also for designing clinical trials and biological experiments, including selection of samples for stem cell research.

LD Adjustment by Multivariate Regression

The effect size of the association of a single SNP with a trait can easily be calculated as the regression coefficient from case-control data. However, when the effect sizes of a number of SNPs are to be combined, for example, in calculating a polygenic risk score or set-based significance, the effect sizes of SNP markers in LD may be given undue weight, as the single-SNP regression coefficients partly replicate the association of the correlated markers. In the extreme case of 2 identical markers, the effect size could be counted double. Therefore some adjustment of single-SNP effect sizes before combining them into summary indicators is desirable.

If, instead of single-SNP regression, multivariate regression over all SNPs is performed, then the correlation between the SNPs is automatically taken into account and eliminated. However, multivariate regression over a large number of variables is numerically unstable and prone to convergence issues and therefore not practically feasible. Moreover, it is impossible when only single-SNP effect sizes are available, for example, from a GWAS, but the original case and control genotypes are not accessible. Nevertheless, the joint regression coefficients can be calculated from the single-SNP regression coefficients if the SNP-SNP correlation is known. Indeed, denoting the *M* vector of single-SNP regression coefficients for *M* SNPs, calculated from standardised genotypic data, by $\tilde{\beta}$, the corresponding vector of joint regression coefficients by $\hat{\beta}$ and the SNP-SNP (LD) correlation matrix of the genotypic data by *D*, it is well known that the ordinary least squares (OLS) estimate for the joint regression is given by the formula

$$\hat{\beta} = D^{-1}\tilde{\beta} \tag{1}$$

(see, e.g., Yang et al. [8]; note that, as did Vilhjálmson et al. [4], we here consider linear regression for simplicity).

To illustrate the effect of the adjustment (equation 1) on the effect sizes of correlated markers, consider the simple case of only 2 markers in correlation *r*, but independent of other markers. In this case

$$D = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad D^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}.$$

The action of this matrix on the 2-vector $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2)^T$ is, by formula (1), the adjustment $\hat{\beta}_1 = (\tilde{\beta}_1 - r\tilde{\beta}_2)/(1 - r^2)$, $\hat{\beta}_2 = (\tilde{\beta}_2 - r\tilde{\beta}_1)/(1 - r^2)$. Its effect can be more clearly seen in the sum and difference of the coefficients,

$$\hat{\beta}_1 + \hat{\beta}_2 = \frac{\tilde{\beta}_1 + \tilde{\beta}_2}{1+r}, \qquad \hat{\beta}_1 - \hat{\beta}_2 = \frac{\tilde{\beta}_1 - \tilde{\beta}_2}{1-r}.$$

The adjustment of the sum (or average) is an intuitive correction for LD, as the weight of 2 highly correlated markers ($r \approx 1$) is reduced to approximately its half. (In the more general case of a block of M_l markers in constant LD with correlation r, the correction factor of the average of the effect sizes will be $1/(1 + [M_l - 1]r)$, so if the correlation is close to 1, then the average effect size of SNPs in the block is divided by approximately the number of SNPs in the block). The adjustment of the difference of the effect sizes is perhaps less intuitive and more problematic, in particular if r is close to –1. If 2 SNPs are in high LD, we expect the single-SNP effect sizes to be very similar, too, so any difference between them will be very important and is hence amplified in the adjustment. For r close to 1, the difference will be amplified very strongly. This can be very problematic when we take proxies for the correlation coefficients from a different sample from the one used for regression, because random fluctuations in the effect sizes between the samples will be amplified if the markers are in high LD. For example, consider 2 SNP markers in LD with correlation $r = 0.98$ and single-SNP effect sizes $\tilde{\beta}_1 = 0.1$, $\tilde{\beta}_2 = 0.09$. Then the adjusted effect sizes will have $\hat{\beta}_1 + \hat{\beta}_2 = 0.096$ and $\hat{\beta}_1 - \hat{\beta}_2 = 0.5$, giving $\hat{\beta}_1 = 0.298$ and $\hat{\beta}_2 = -0.202$. Clearly the sum of the effect sizes has been approximately halved, but their small difference is strongly amplified, leading in particular to a severe adjustment of both effect sizes. If furthermore the LD is taken from a proxy with, for example, $r = 0.99$, the resulting adjusted effect sizes have sum $\hat{\beta}_1 + \hat{\beta}_2 = 0.0955$ and difference $\hat{\beta}_1 - \hat{\beta}_2 = 1$, giving $\hat{\beta}_1 = 0.54775$ and $\hat{\beta}_2 = -0.452$. While the adjusted sum is almost the same as before, we here see an even more severe adjustment of the difference and therefore of both effect sizes, showing the instability of the straightforward OLS adjustment. Note that when the markers are highly negatively correlated, so r is close to –1, then the sum and difference of the effect sizes swap roles in the above discussion.

Tikhonov-Regularised LD Adjustment

Generally, in the presence of high LD, the correlation matrix D will be numerically ill-conditioned or even singular, and therefore the calculation of its inverse in equation 1 will be numerically difficult or unstable, and will in any case lead to an adjustment matrix which has the tendency to amplify differences between the effect sizes of correlated markers, even when these differences are due to random fluctuations between different samples.

A standard method of avoiding such difficulties arising from an ill-conditioned matrix is Tikhonov regularisation [9, sect. 5.3]. (In a regression context, this method is associated with ridge regression, but we prefer to consider the more general concept of regularising the inverse problem (equation 1). The idea here is to make the non-negative definite correlation matrix strictly positive definite, with a positive lower bound, by adding a multiple of the unit matrix. This gives a regularised adjustment of the effect sizes of the general form

$$\beta_{\text{adj}} = (D + TI)^{-1}\tilde{\beta}, \tag{2}$$

where $T > 0$ is the regularisation constant. The operator norm of the inverse matrix in equation 2 is bounded by $1/T$, ensuring that there is no uncontrolled amplification of effect sizes β. Note that the operator norm of a matrix is the maximal factor of amplification of a vector through multiplication with this matrix. In the present situation, where D is a correlation matrix, Tikhonov regularisation can be interpreted in the following way. Separating the diagonal entries 1 of the matrix D from the correlation coefficients of SNP pairs by writing $D = \hat{D} + I$, where $\hat{D}$ is the reduced matrix of correlation coefficients where the diagonal entries are set to 0, we can rewrite the adjustment formula 2 equivalently in the form

$$\beta_{\text{adj}} = \left(\hat{D} + (1+T)I\right)^{-1}\tilde{\beta} = \frac{1}{1+T}\left(\frac{\hat{D}}{1+T} + I\right)^{-1}\tilde{\beta}. \tag{3}$$

This shows that, apart from an overall division by the constant $1 + T$, the regularisation consists of dividing each pairwise correlation coefficient by the number $1 + T$. Thus, the regularisation avoids the singularities at correlation $r = \pm 1$ by just scaling down all the correlation coefficients. This means that instead of adjusting for the LD present in the markers, the regularised adjustment 3 only adjusts for a fixed fraction of this LD in each pair of markers. The effect sizes for a pair of markers in LD with some value of r^2 are only corrected as if the LD were $r^2/(1 + T)^2$.

In consequence, T needs to be fairly small for effective LD adjustment to occur; for large values of T, the values β_{adj} will essentially be uniformly scaled down values of $\tilde{\beta}$, rendering the adjustment ineffective when the β_{adj} are used in a subsequent analysis that introduces its own scale, such as linear or logistic regression.

Although this downscaling is applied uniformly to all correlation coefficients, the effect on the adjustment of effect sizes will depend on the size of the LD blocks. Due to the LD structure of the genome, the SNP-SNP correlation matrix D is approximately block-diagonal, that is, it is composed of square matrices D_l of noticeable correlation strung along the diagonal of D, with small entries outside these blocks [10]. If we neglect these small entries, the regularisation and inverse in formula 3 apply to each LD block matrix D_l separately. While the operator norm of the regulariser $(1 + T)I$ is equal to $1 + T$ independently of the block size, the norm of $\hat{D}_l$, the part of $\hat{D}$ restricted to the block, will depend on the correlation coefficients in the block, but even in case of extreme LD is bounded by $M_l - 1$, where M_l is the number of SNPs in the block (see Appendix B). If $1 + T$ is larger than the norm of $\hat{D}_l$, then, apart from the overall scaling factor, the adjustment of each effect size will only be a minor correction.

As an illustrative, albeit somewhat cartoonish example, suppose the LD matrix is exactly of block-diagonal form, and the component block D_l has constant correlation r_l, so the entries of D_l are 1 on the diagonal and r_l everywhere else. Then the adjustment (equation 3) can be calculated in explicit form (see Appendix B), giving for each $\tilde{\beta}_{\text{j}}$ in the corresponding LD block

$$\beta_{\text{adj},j} = \frac{1}{1 - r_l + T}\left(\tilde{\beta}_j - \frac{M_l r_l}{1 - r_l + T + M_l r_l}\left\langle \tilde{\beta} \right\rangle_l\right), \tag{4}$$

where $\langle \tilde{\beta} \rangle_l$ is the mean of all observed effect sizes in the LD block. We can see that the adjustment consists of a scaling with factor $1/1 - r_l + T$ and a constant shift, the latter dependent on the average of SNP effect sizes in the LD block. For large T, the dependence of the scaling factor on the correlation r_l, which lies between –1 and 1, is of minor significance, and the scaling factor is approximately equal to the common overall factor $(1 + T)^{-1}$ of equation 3. The shift term depends on the average effect size, but will be small unless the product $M_l r_l$ is of noticeable size compared to T. For larger values of T, this will only be the case for large blocks of high LD.

As a further illustration of this effect, we show the result of a simulated example in Figure 1. Genotypes for 10,000 cases and 10,000 controls in 100 SNPs were randomly generated. The first block, comprising 6 SNPs, has an odds ratio of 1.1 and LD $r^2 \approx 0.9$ between SNPs, the second block, comprising 50 SNPs, has an odds ratio of 1.1 and LD $r^2 \approx 0.4$ between SNPs. The remaining 54 SNPs were generated unassociated and without LD, as a baseline for comparison. Figure 1a shows the SNP-SNP correlation matrix; there is some small random correlation between the LD blocks and the baseline SNPs. The vector of single-SNP effect sizes, calculated as logistic regression coefficients, was then subjected to regularised LD adjustment according to equation 2, with regularisation parameter $T = 0.01$, $T = 1$ and $T = 10$. The comparison of the resulting adjusted effect sizes with the raw (single-SNP) values is shown in Figure 1b–d. For mild regularisation ($T = 0.01$), the LD blocks show adjustment according to the size of correlation, with some strong corrections in the small high-LD block. For moderate regularisation ($T = 1$), the outcome is similar, but extreme adjustment is avoided, and the correction factors (line slopes) are more aligned. With a stronger regularisation ($T = 10$), the large block of moderate LD still receives adjustment, but the small block of high LD tends to fall in line with the uncorrelated baseline SNPs that do not need adjustment; as the regularisation constant exceeds the block size, the regularisation makes the adjustment less effective.

The size of the LD blocks in the genome varies greatly. In Reich et al. [11], blocks of significant LD ranged in size from 6 to 155 kb, with an average of about 60 kb in North American individuals of European descent. Some LD blocks can be very large, for example, a block in the MHC region on chromosome 6 spans up to 5 Mb [12]. The relevant figure for the present considerations is the number of SNPs in an LD block. As an example, a typical set of summary statistics, including imputed SNPs after quality controls, contains about 10 million SNPs. We estimated about 4 SNP/kb on average in Alzheimer's disease summary statistics [13]. The Hapmap3 list of SNPs used in LDscore regression [14] contains about 1.2 million SNPs, giving an average of 0.5 SNP/kb. The publicly available 1,000 genomes data, often used as a reference panel for LD calculation, contains about 2.3 millions SNPs for a European population, giving an estimated 1 SNP/kb. These data show that the size of LD blocks can be expected to range from about 3 to 600 SNPs or, in some cases, even more.

When setting the regularisation constant T explicitly, one would naturally choose a small value (in particular, smaller than the smallest number of SNPs in an LD block) in order to keep effective LD adjustment; however, the constant T may turn out to be large when it arises implicitly and out of the researcher's direct control, as shown in the following section.

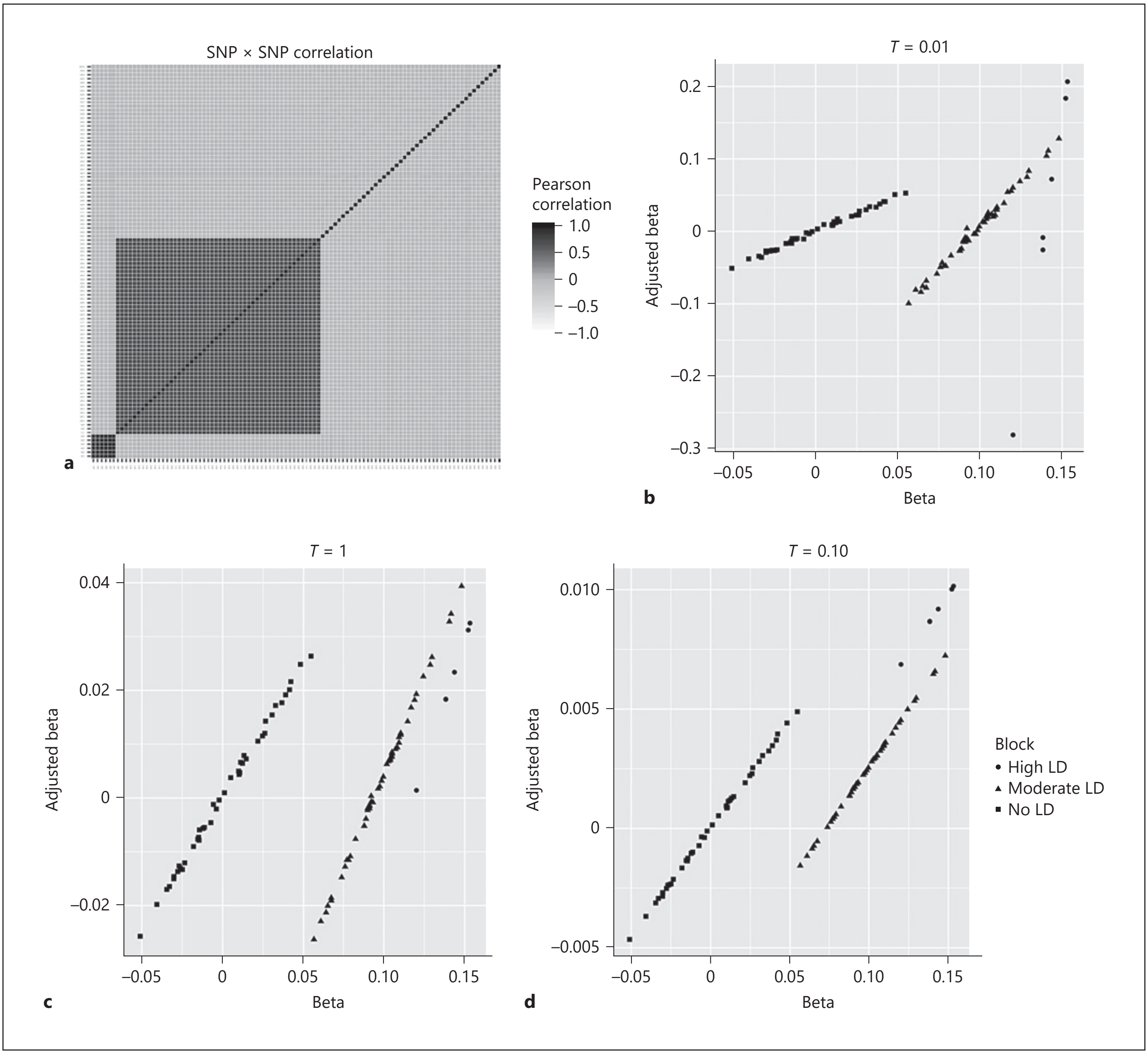

Fig. 1. Regularised LD adjustment for 100 simulated SNPs (10,000 cases and 10,000 controls). **a** SNP-SNP correlation matrix. **b–d** Comparison of raw (single-SNP) effect sizes (beta) with adjusted effect sizes (adjusted beta) for regularisation parameter $T = 0.01$, $T = 1$ and $T = 10$.

LDpred's Bayesian Approach in the Context of Regularised LD Adjustment

The fundamental idea of LDpred, as explained in Vilhjálmsson et al. [4], is to correct the vector of single-SNP effect sizes or marginal least-squares effects $\tilde{\beta}$ by replacing them with the expectation values of a Bayesian update of an informationless prior, using these effect sizes as empirical input. In the infinitesimal model (LDpred-inf), it is assumed that each SNP is a priori equally likely to have an effect. This is modelled by setting the prior distribution to i.i.d. normal with mean 0 and variance h^2/M, where h^2 is the total expected variance (disease heritability) and M is the total number of SNPs. Given any supposed vector of real effect sizes β, the conditional probability distribution of observed single-SNP effect sizes $\tilde{\beta}$ is calculated from the regression model. Combined with the prior distribution, this gives the joint distribution of β and $\tilde{\beta}$. By Bayesian inversion, a posterior distribution is calculated from the observed effect sizes $\tilde{\beta}$. The posterior distribution is multivariate normal with parameters (mean and covariance matrix)

$$E\left(\beta\,|\,\tilde{\beta}\right)=\left(D+\frac{M\left(1-h_l^2\right)}{h^2N}I_M\right)^{-1}\tilde{\beta}, \tag{5}$$

$$\operatorname{Var}\left(\beta\,|\,\tilde{\beta}\right)=\left(\frac{N}{1-h_l^2}D+\frac{M}{h^2}I_M\right)^{-1}. \tag{6}$$

Here N is the number of individuals in the data set from which the observed effect sizes were calculated, D is the SNP-SNP correlation matrix estimated from the data set, $1 - h_l^2$ is the variance of the residuals per individual in the regression model, and I_M is the $M \times M$ unit matrix. We show the derivation of these formulae in Appendix A, both for ease of reference and because the corresponding passage in Vilhjálmsson et al. [4] contains a number of misprints, including a spurious factor of $1 - h_l^2$ in the formula for the expectation $E(\beta|\tilde{\beta})$ on p. 587.

The expected vector of effect sizes from the posterior distribution (equation 5) is then taken to be a Bayesian LD-adjusted vector of effect sizes. Assuming that the adjustment is applied to LD regions of small size compared to the total number M of SNPs, the approximation $1 - h_l^2 \approx 1$ is applied [4, pp. 578–579]. Clearly, the adjustment is then just a Tikhonov-regularised LD correction as in equation 2 with the region-independent constant

$$T=\frac{M}{h^2N}. \tag{7}$$

In particular, the considerations at the end of section 2 above about the dependence of the effect of regularisation on the size of the LD region apply.

Thus, the main contribution of LDpred is to provide a rationale, based on the Bayesian model, for a particular choice of the regularisation constant T. Taken at face value, LDpred's regularisation constant (equation 7) is determined by the total number of SNP M, the sample size N and the estimated overall heritability h^2 – in fact, the derivation shows that the estimated heritability per SNP marker h^2/M is the basic quantity here. For a large full-genome study, M will be on the order of 10^6–10^7 and the number of individuals genotyped is typically in the range $N \approx 10^4$–10^5. If the heritability h^2 lies below 1, this will give a regularisation constant T of an order of 10^2 or higher, which is fairly large and exceeds the size of the smaller LD blocks.

We note that the ostensible dependence of T on M and N in equations 7 and 5 is somewhat misleading. Indeed, in LDpred the heritability h^2 is estimated from the data using the formula

$$h^2=\frac{\left(\overline{\chi^2}-1\right)M}{\bar{l}N}$$

(see Vilhjálmsson et al. [4, p. 579] and Bulik-Sullivan et al. [15, supplementary note]), where $\overline{\chi^2}$ is the average per-SNP χ^2 statistic taken over all SNPs and $\bar{l}$ is the average LD score taken over all SNPs. When this estimate is used in equations 7 and 5, the numbers M and N cancel out and play no immediate role in determining the regularisation constant,

$$T=\frac{\bar{l}}{\overline{\chi^2}-1}. \tag{8}$$

As a result, the effect of the adjustment (equation 3) on each LD block will be determined by the trade-off between the operator norm of the reduced correlation matrix for this block, an average LD score and the average χ^2 statistic across the markers. The LD score of the j-th marker is approximately

$$l_j=\sum_k r_{jk}^2,$$

summing over the neighbouring markers in a window of size K. Estimating the average value of r^2 in this window at $r^2 = 0.2$ for $K = 10$ and $r^2 = 0.1$ for $K = 100$ and taking a typical value of $\overline{\chi^2} = 1.1$ (e.g., $\overline{\chi^2} = 1.0783$ in Kunkle et al. [13]), we find $T = 20$ for $K = 10$ and $T = 100$ for $K = 100$. Again, these values are so large that the regularisation will render the LD adjustment ineffective in smaller LD blocks.

There is a non-infinitesimal variant of LDpred based on a Bayesian prior where the normal distribution for each SNP is mixed with a delta distribution at 0, intended to model a situation where a certain fraction of genetic variants is not at all causal. This provides a parameter for optimisation but makes the analysis much less transparent and in particular does not allow an explicit formula for the posterior distribution beyond observation of per-SNP scaling with the constant $1 + T$ with T as in equation 7, see Vilhjálmsson et al. [4, p. 588]. The LD adjustment is not made explicit, and Márquez-Luna et al. [7] recommend exclusion of SNPs from long-range LD regions. For these reasons, we do not discuss non-infinitesimal LDpred further here.

The PRS-CS method [5] and the SBayesR method [6] use a more sophisticated prior composed of a mixture of normal distributions with optimised parameters. These approaches are shown to achieve higher prediction accuracy than LDpred, but also preclude explicit analysis of the resulting posterior distribution.

Variants of Regularisation

More recently, a functionally informed variant of the infinitesimal model for LDpred has been proposed, see "LDpred-funct-inf" in Márquez-Luna et al. [7]. Here the multiple of the unit matrix

$$\frac{M}{h^2} I_M$$

in equation 6 is replaced with the non-constant diagonal matrix diag($1/[c\sigma_1^2],\ldots,1/[c\sigma_{M+}^2]$) where $\sigma_i^2 > 0$ is an estimated per-SNP heritability for the i-th SNP,

$$c = h^2 / \sum_{i=1}^{M_+} \sigma_i^2$$

is a normalisation constant and the number of SNPs has been reduced to M_+ by removing all SNPs which do not have a positive σ_i^2. This gives (see Appendix A)

$$\beta_{\text{adj}} = \left(D + \text{diag}\left(\frac{1}{Nc\sigma_i^2} \right) \right)^{-1} \tilde{\beta}. \qquad (9)$$

Using the matrix $\widehat{D}$ defined above and abbreviating $\gamma_i = 1 + 1/(Nc\sigma_i^2)$ for the i-th marker, we observe that the matrix in equation 9 can be rewritten as

$$(\widehat{D} + \text{diag}(\gamma_i))^{-1} = \text{diag}(1/\surd\,\gamma_i)\widetilde{D}^{-1}\,\text{diag}(1/\surd\,\gamma_i)$$

where $\widetilde{D}$ is a correlation matrix with non-diagonal entries

$$\tilde{D}_{jk} = \frac{r_{jk}}{\sqrt{\gamma_j}\sqrt{\gamma_k}}.$$

Thus, in comparison with the plain adjustment (equation 1), the correlation coefficients are scaled down and the effect sizes are also directly reduced. In order to assess the extent of this scaling, we note that the mean of $Nc\sigma_i^2$ is equal to $Nh^2/M_+ = 1/T$ with T as in equation 7, but where M is replaced by the generally smaller number M_+ and h_I^2 is as before neglected. By Jensen's inequality,

$$\text{mean}\,\gamma_i \geq 1 + \frac{1}{\text{mean}\left(Nc\sigma_i^2\right)} = 1 + T,$$

so if T is big, then γ_i will be large for a large proportion of SNPs. For these SNPs, there will be a strong direct downscaling of the effect sizes with a factor varying between SNPs in dependence on their estimated σ_i^2, but the SNP-SNP correlation coefficient measuring the actual LD will also be scaled down severely. The effect on these SNPs, then, will be a per-SNP scaling rather than an LD adjustment. For SNPs with a large per-SNP heritability σ_i^2, the corresponding γ_i may be smaller and close to the minimal possible value 1, in which case the adjustment (equation 9) will be close to the straightforward OLS estimate (equation 1) with little direct downscaling.

In the overall result, a PRS formed with the adjusted effect sizes will give great prominence to SNPs with a high previously estimated per-SNP heritability, with LD adjustment essentially restricted to these markers.

A rather different approach to the question of LD adjustment for the purpose of calculation of PRS was proposed in Baker et al. [3]. The POLARIS method uses, instead of equation 2, the adjustment

$$\beta_{\text{adj}} = \left(D + TI\right)^{-\frac{1}{2}} \tilde{\beta},$$

either without regularisation ($T = 0$) or very mildly regularised ($T \ll 1$). The rationale behind this is that the OLS formula 1 eliminates all correlations on the right-hand side $\beta^T X$ of the linear regression formula, so it adjusts for correlation both in the coefficients β and the data vectors X. However, although the effect sizes $\widetilde{\beta}$ are obtained as single-SNP regression coefficients, they enter the PRS as per-SNP effect sizes which may arise from different samples and can be connected functionally as well as by LD. Hence, forming a PRS from a genotype vector x in the form $(D^{-1}\widetilde{\beta})^T x = \widetilde{\beta}^T D^{-1}x$ can be considered an overcorrection, and the form $\widetilde{\beta}^T D^{-1/2}x$ (or a mildly regularised form of this) is preferred, which can be taken as a correction of x for LD with the actual data correlation matrix D while taking the single-SNP effect sizes at face value. This form of LD adjustment for PRS was shown to be effective [3].

From the practical point of view, the matrix $D^{1/2}$, if not directly singular, is considerably better conditioned than the matrix D, as the square root function moves the very small eigenvalues disproportionately farther away from 0. In this sense, taking the square root is a form of regularisation which runs less risk of obscuring the correlation structure encoded in the LD matrix than Tikhonov regularisation with a large constant T.

Discussion

In order to avoid spurious contribution from associated SNPs in medium to high LD, it is essential to take the LD between SNPs into account when preparing SNP effect sizes for inclusion in aggregates for further analysis such as in PRS or (converted to *p* values) in gene sets as in Brown's method (MAGMA [1]). When assessing different methods of LD adjustment, a main concern may be the prediction accuracy achieved in the end; however, since the ultimate aim of polygenic statistical analysis is not just to find a suitable measure of disease risk, but to interpret and understand its structure and thus to identify the co-causal genes, we contend that transparency of the choices and calculations involved is a further crucial characteristic. A comparison of prediction accuracy for different approaches can be found in Ge et al. [5] and Lloyd-Jones et al. [6]; in the present discussion we focus on the interpretability and possible misconceptions on the side of the user.

The "P + T" approach avoids the issue by using independent (low LD) SNPs only, but this may lead to loss of information and hence of prediction accuracy. Nevertheless, this method is fast, transparent and easy to interpret; for example, the identification of SNPs of which an individual at high risk for the disease has 2 risk alleles can inform which variants or genes are mostly contributing to the PRS.

Direct multivariate regression of case-control data would automatically account for correlation between SNPs, but is not practically feasible due to computational problems when the number of SNPs included is large, and also due to the fact that usually only single-SNP regression coefficients are available from previous studies. It can be replaced computationally by using a proxy SNP-SNP correlation matrix in the OLS formula (equation 1); the computation of the inverse matrix, however, usually requires some regularisation such as the standard Tikhonov regularisation (equation 2). Normally, one would expect to use a fairly small regularisation parameter *T*, as strong regularisation is essentially equivalent to an artificial downscaling of the LD correlation coefficients. Our analysis indicates that *T* should not exceed 1 in order to ensure effective LD adjustment even of small blocks of highly correlated SNPs.

LDpred appears to offer a different, more data-driven way, as it is based on a Bayesian model involving quantities estimated from the available data themselves. However, on inspection, LDpred with the infinitesimal prior turns out to give a regularised OLS estimate for multivariate regression coefficients. The contribution of the Bayesian framework is that it provides a specific choice of the regularisation parameter. This saves the user the trouble of choosing a suitable regularisation parameter or of even realising that regularisation is taking place, but there are 2 caveats.

First, the choice of the regularisation parameter is not very transparent; although it ostensibly seems to depend on the study parameters of sample size and number of SNPs, the implicit heritability estimate used in fact removes the direct dependence on these manifest numbers and instead determines the regularisation parameter in terms of average LD scores and the average χ^2 statistic. Therefore the resulting LD adjustment will be non-local; to what extent the effect sizes of SNPs in some region will be corrected depends not only on the LD between the SNPs, but also on the effect sizes and LD found on average throughout the genome.

Second, the regularisation parameter resulting from the Bayesian estimate may be rather large. In this case, adjustment will only be made for a possibly small fraction of the actual LD. This effect is particularly pronounced in small LD blocks, which even in case of extreme LD receive no more LD adjustment than if they were uncorrelated. Note that the equal downscaling of all effect sizes has no effect when the effect sizes are subsequently linearly combined into a single regression variable. This gives rise to a situation where, out of the control or knowledge of the user and contrary to the expectation raised by its name, LDpred does not perform full LD adjustment of SNPs, even in high LD and possibly with large effect sizes, resulting in the spurious contributions, for example, to PRS that accompany the use of uncorrected single-SNP effect sizes.

The functionally informed variant of LDpred varies the strength of regularisation for each SNP based on its estimated per-SNP heritability σ_i^2. As a result, SNPs with low or medium σ_i^2 will have their effect sizes directly scaled down, but not adjusted for LD. SNPs with high σ_i^2 will be promoted due to less penalisation of their effect sizes and adjusted for LD among each other. Thus, LDpred-funct-inf can be considered as a fuzzy version of P + T or, if the SNPs with high σ_i^2 are in LD, of fuzzy pruning combined with a mildly regularised OLS estimate. Therefore LDpred-funct-inf points in the direction of what the field wants to achieve by intelligent pruning [16]. In comparison to LDpred-inf, which imposes a uniformly regularised, possibly much diluted LD adjustment on all SNPs on the basis of a Bayesian model, LDpred-funct-inf aims more directly at a focus on SNPs that appear likely to have explanatory power and subjects their effect sizes to LD correction with little adjustment while eliminating or diminishing the effect sizes of the other SNPs. This method achieves higher prediction accuracy than LDpred-inf by using the estimated per-SNP heritabilities, but, as the selection of SNPs for emphasis remains internal, is not as transparent as a straightforward method of pruning and mildly regularised OLS. This also applies to methods with variant Bayesian priors, such as non-infinitesimal LDpred, PRS-CS and SBayesR. These methods may reach higher prediction accuracy, likely due to promoting the most associated SNPs by reduced and more relaxed LD adjustment. This may make practical sense, but ultimately transcends the paradigm of approximating the "true" effect sizes one would obtain from multivariate regression.

The POLARIS method uses the matrix square root for a regularisation that runs less risk of obscuring the LD structure encoded in the correlation matrix than Tikhonov regularisation with a large parameter, but at additional computational cost.

In conclusion, faced with an array of automated methods of processing single-SNP effect sizes for LD, some of considerable complexity, but not always guaranteed to perform the expected task of LD adjustment, the user may consider whether, beyond mere prediction accuracy, the aim of identifying the polygenic elements of a disease is not better served by a simple and transparent method of cautious pruning and gently regularised OLS adjustment for LD.

Acknowledgement

We thank the referee for helpful and constructive comments.

Statement of Ethics

This paper is exempt from Ethical Committee approval as a theoretical study that did not collect or use any real data.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This study was supported by Dementia Research Institute (UK-DRI-3003), Dementias Platform UK (MR/L023784/2) and MRC Centre for Neuropsychiatric Genetics and Genomics (MR/L010305/1).

Author Contributions

Both authors contributed equally to conducting the research and writing the paper.

Appendix A

Derivation of the LDpred Adjustment Formula

In the following we summarise the calculation and assumptions from Vilhjálmson et al. [4], that lead to equations 5 and 6, considering the infinitesimal model with LD only. Let X be the genotype matrix, with M columns (corresponding to the number of SNPs) and N rows (corresponding to the number of individuals genotyped), and let Y be the N vector of phenotypes. It is assumed that Y and the columns of X are standardised to mean 0 and variance 1. The linear regression model is

$$Y = X\beta + \varepsilon, \tag{10}$$

where the errors ε are assumed to be independent and identically normally distributed with expectation $E\varepsilon = 0$ and covariance matrix $E\varepsilon\varepsilon^T = (1-h_l^2)I_N$.

The OLS estimate for the regression coefficients is

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = D^{-1}\frac{X^T Y}{N},$$

using in the latter formula the convention of Vilhjálmsson et al. [4], where D is an abbreviation for

$$D = \frac{X^T X}{N}.$$

If the columns of the data matrix X are treated as orthogonal (independent), one obtains the estimate

$$\tilde{\beta} = \left(NI_M\right)^{-1} X^T Y = \frac{X^T Y}{N},$$

and this is equal to the result of separate linear regression on each SNP.

Now assume the "true" effect sizes are β, so equation 10 holds, where only Y and ε are random variables. Under this assumption of true effects β, the distribution parameters of $\widetilde{\beta}$ can be calculated as follows:

$$E\left(\tilde{\beta}\,|\,\beta\right) = E\left(\frac{X^T Y}{N}\right) = E\left(\frac{X^T(X\beta+\varepsilon)}{N}\right) = \frac{X^T X}{N}\beta + \frac{X^T}{N}E\varepsilon = D\beta \tag{11}$$

for the expectation and

$$\begin{aligned}\mathrm{Var}\left(\tilde{\beta}\,|\,\beta\right) &= E\left(\tilde{\beta}\tilde{\beta}^T\,|\,\beta\right) - E\left(\tilde{\beta}\,|\,\beta\right)E\left(\tilde{\beta}\,|\,\beta\right)^T\\ &= E\left(\frac{X^T Y}{N}\frac{X^T Y}{N}^T\right) - D\beta\left(D\beta\right)^T\\ &= E\left(\frac{X^T\left(X\beta+\varepsilon\right)\left(X\beta+\varepsilon\right)^T X}{N^2}\right) - D\beta\beta^T D\\ &= E\left(\frac{X^T X}{N}\beta\beta^T\frac{X^T X}{N}\right) + \frac{1}{N^2}X^T\left(E\varepsilon\right)\beta^T X^T X + \frac{1}{N^2}X^T X\beta\left(E\varepsilon\right)^T X\\ &+ E\left(\frac{X^T\varepsilon\varepsilon^T X}{N^2}\right) - D\beta\beta^T D\\ &= E\left(\frac{X^T\varepsilon\varepsilon^T X}{N^2}\right) = \frac{1}{N^2}X^T\left(E\varepsilon\varepsilon^T\right)X = \frac{1-h_l^2}{N^2}X^T X = \frac{1-h_l^2}{N}D \end{aligned} \tag{12}$$

for the covariance matrix. From equations 11 and 12 we then obtain the conditional probability density

$$p\left(\tilde{\beta} \mid \beta\right) = \text{const}\exp\left(-\frac{\left(\tilde{\beta} - D\beta\right)^T D^{-1}\left(\tilde{\beta} - D\beta\right)}{2\left(1-h_l^2\right)/N}\right)$$

and further use the assumed density for a multivariate Gaussian prior with covariance matrix C_p

$$p\left(\beta\right) = \text{const}\exp\left(-\frac{1}{2}\beta^T C_p^{-1}\beta\right)$$

to set up the joint density

$$p\left(\tilde{\beta},\beta\right) = p\left(\tilde{\beta} \mid \beta\right)p\left(\beta\right) = \text{const}\exp\left(-\frac{1}{2}\left(\frac{\left(\tilde{\beta} - D\beta\right)^T D^{-1}\left(\tilde{\beta} - D\beta\right)}{\left(1-h_l^2\right)/N} + \beta^T C_p^{-1}\beta\right)\right).$$

We now rewrite the term in the exponential as a quadratic expression in β, that is, to be of the form $(\beta - m)^T \Sigma^{-1}(\beta - m) + \mathit{const}$, where the constant does not depend on β and m is subsequently interpreted as posterior mean, Σ as posterior covariance matrix for β. Since

$$\begin{aligned}
&\frac{\left(\tilde{\beta} - D\beta\right)^T D^{-1}\left(\tilde{\beta} - D\beta\right)}{\left(1-h_l^2\right)/N} + \beta^T C_p^{-1}\beta \\
&= \frac{\tilde{\beta}^T D^{-1}\tilde{\beta}}{\left(1-h_l^2\right)/N} - \frac{\beta^T\tilde{\beta}}{\left(1-h_l^2\right)/N} - \frac{\tilde{\beta}^T\beta}{\left(1-h_l^2\right)/N} + \frac{\beta^T D\beta}{\left(1-h_l^2\right)/N} + \beta^T C_p^{-1}\beta \\
&= \beta^T\left(\frac{D}{\left(1-h_l^2\right)/N} + C_p^{-1}\right)\beta - \frac{\beta^T\tilde{\beta}}{\left(1-h_l^2\right)/N} - \frac{\tilde{\beta}^T\beta}{\left(1-h_l^2\right)/N} + \text{const},
\end{aligned}$$

one can read off

$$\Sigma^{-1} = \frac{D}{\left(1-h_l^2\right)/N} + C_p^{-1} = \frac{N}{1-h_l^2}D + C_p^{-1},$$

and

$$\frac{N}{1-h_l^2}\tilde{\beta} = \Sigma^{-1}m = \left(\frac{N}{1-h_l^2}D + C_p^{-1}\right)m,$$

so

$$m = \left(D + \frac{1-h_l^2}{N}C_p^{-1}\right)^{-1}\tilde{\beta}.$$

When using the infinitesimal non-informative prior where $C_p = h^2/MI_M$, this gives the formulae 5 and 6. For the functionally informed prior with previously estimated per-SNP heritability σ_i^2 for the i-th SNP and $c = h^2/\sum_i\sigma_i^2$, we have the diagonal covariance matrix $C_p = \text{diag}(\sigma_1^2, \ldots, \sigma^2{}_{\text{M+}})$ with reduced dimension M_+ due to omission of SNPs which do not have $\sigma_i^2 > 0$, and the adjustment formula 9 follows when h_l^2 is neglected.

Appendix B

The Importance of LD Block Size

Consider an LD block of size M_l and denote by D_l the corresponding part of the SNP-SNP correlation matrix. We write $D_l = I + \widehat{D}_l$, where I is the $M_l \times M_l$ unit matrix and $\widehat{D}_l$ is the reduced correlation matrix with entries r_{jk} off-diagonal and zeros on the diagonal. With respect to the vector norm

$$\| x \|_p = \left(\sum_{j=1}^{M_l} | x_j |^p\right)^{\frac{1}{p}},$$

where $p \geq 1$, the operator norm of $\widehat{D}_l$ is the maximum value of the ratio $\|\widehat{D}x\|_p/\|x\|_p$ as x ranges over all M_l vectors; since

$$\begin{aligned}
\| \hat{D}x \|_p &= \left(\sum_j \left|\sum_{k\neq j} r_{jk}x_k\right|^p\right)^{\frac{1}{p}} \leq \left(\sum_j \left(\sum_{k\neq j} | r_{jk} |^q\right)^{\frac{p}{q}}\left(\sum_{k\neq j} | x_k |^p\right)\right)^{\frac{1}{p}} \\
&\leq \left(M_l - 1\right)^{\frac{1}{q}}\left(\sum_k\sum_{j\neq k} | x_k |^p\right)^{\frac{1}{p}} = \left(M_l - 1\right)^{\frac{1}{q}}\left(M_l - 1\right)^{\frac{1}{p}} \| x \|_p
\end{aligned}$$

(where $1/p + 1/q = 1$), we see that the operator norm of $\widehat{D}_l$ cannot exceed $M_l - 1$.

Now suppose the correlation in the block is constant, so $r_{jk} = r_l$ with fixed r_l, and neglect the correlations with markers outside the block. Then, considering SNPs in the corresponding LD block only, the relationship between the observed effect size $\tilde{\beta}_j$ and the adjusted effect size $\beta_{\text{adj},j}$ by equation 3 is

$$\tilde{\beta}_j = (1 - r_l + T)\beta_{\text{adj},j} + r_l M_l <\beta_{\text{adj}}>_l,$$

where $<\ldots>_l$ denotes averaging over the LD block of length M_l. Hence

$$<\tilde{\beta}>_l = (1 - r_l + M_l r_l + T) <\beta_{\text{adj}}>_l,$$

and equation 4 follows by substitution and rearrangement.

References

1 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLOS Comput Biol. 2015 Apr; 11(4):e1004219.

2 Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al.; International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009 Aug;460(7256):748–52.

3 Baker E, Schmidt KM, Sims R, O'Donovan MC, Williams J, Holmans P, et al. POLARIS: polygenic LD-adjusted risk score approach for set-based analysis of GWAS data. Genet Epidemiol. 2018 Jun;42(4):366–77.

4 Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015 Oct;97(4):576–92.

5 Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr;10(1):1776.

6 Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun. 2019 Nov;10(1):5086.

7 Márquez-Luna C, Gazal S, Loh PR, Kim SS, Furlotte N, Auton A. 23andMe Research Team & Price A.L. (2019). Modeling Functional Enrichment Improves Polygenic Prediction Accuracy in UK Biobank and 23andMe Data Sets. bioRxiv; DOI: 10.1101/375337.

8 Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012 Mar;44(4):369–75.

9 Kress R. Numerical Analysis. New York: Springer; 1998.

10 Slatkin M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun; 9(6):477–85.

11 Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001 May; 411(6834):199–204.

12 de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006 Oct;38(10): 1166–72.

13 Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al.; Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet. 2019 Mar;51(3): 414–30.

14 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015 Nov;47(11):1236–41.

15 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291–5.

16 Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. Genetics. 2019 Apr;211(4):1131–41.