Title

Polygenic scores using summary statistics via penalized regression

Author list

Timothy Shin Heng Mak, 1

Robert Milan Porsch, 2

Shing Wan Choi, 2

Xueya Zhou, 2

Pak Chung Sham, 1, 2, 3

Affliations

1. Centre for Genomic Sciences, University of Hong Kong

2. Department of Psychiatry, University of Hong Kong

3. State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong

Correspondence

Timothy Shin Heng Mak (`pcsham@hku.hk`)*

Pak Chung Sham (`pcsham@hku.hk`)**

Abstract

Polygenic scores (PGS) summarize the genetic contribution of a person's genotype to a disease or phenotype. They are useful in a wide variety of analyses of genetic data. Many possible ways of calculating polygenic scores have been proposed, and recently there is much interest in methods that incorporate information available in published summary statistics. As there is no inherent information on linkage disequilibrium (LD) in summary statistics, a pertinent question is whether we can make use of LD information available elsewhere to supplement such analyses. To answer this question we proposed a method for constructing PGS using summary statistics and a reference panel in a penalized regression framework, which we called lassosum. We also proposed a general method for choosing the value of the tuning parameter in the absence of validation data. Our simulation results suggested that lassosum is faster and more robust than other similar methods in most scenarios. We also found that accounting for LD with a reference panel is beneficial only when the signals from the data are strong. In the presence of summary statistics from a large number of SNPs, clumping may both enhance or decrease the performance of standard PGS, although its effects on lassosum is attenuated. lassosum combined with pre-filtering by clumping appears to be a robust and reliable option for calculating predictive PGS.

Introduction

A vast number of twin studies as well as recent genome-wide association studies have demonstrated that a large proportion of the variance in liability to common diseases and human traits is due to genetic differences between individuals (Polderman *et al.*, 2015; Yang *et al.*, 2011; Bulik-Sullivan *et al.*, 2015). These studies have also made clear that only a very small proportion of the total genetic contribution can be unambiguously attributed to variation in particular loci of the genome. The vast majority of such genetic contribution is thus spread across the huge landscape of the genome, with many loci each contributing a small, almost undetectable effect on the phenotypes (Dudbridge, 2013, 2016). One important source of evidence towards this conclusion is from studies that examined the association of polygenic predictors of diseases/traits, where it is repeatedly found that SNPs that are not themselves significantly associated with the phenotypes can, by being aggregated as a score, be very significantly associated with the phenotypes, even in totally unrelated samples (Agerbo *et al.*, 2015; Byrne *et al.*, 2014; Evans *et al.*, 2009; Wei *et al.*, 2009; Purcell *et al.*, 2009; Ripke *et al.*, 2013; Speliotes *et al.*, 2010; Machiela *et al.*, 2011; Stahl *et al.*, 2012; Martin *et al.*, 2015; Chang *et al.*, 2015). A particular remarkable demonstration is that persons with such *polygenic scores* for schizophrenia at the top 10 percentile of the population can be at more than 10 times the risk of having the disease than those at the bottom 10 percentile (Ripke *et al.*, 2014; Agerbo *et al.*, 2015), raising hope that one day a person's risk for many common disease can be accurately assessed simply by the examination of one's genome.

Thus, there is considerable interest in the calculation of such polygenic scores (PGS) in GWAS and Genome-wide meta-analyses, where they are also known as risk scores (Ripke *et al.*, 2013; Domingue *et al.*, 2014), polygenic risk scores (e.g. Euesden *et al.*, 2015; Byrne *et al.*, 2014; Agerbo *et al.*, 2015; Dudbridge, 2013), and allelic scores (Burgess and Thompson, 2013; Evans *et al.*, 2013). In a typical application, a unique PGS is assigned to each individual based on the person's genotype. The score summarizes the genetic contribution to a particular disease or phenotype for that indivdiual given his/her genotype. They are then used for testing of complex genetic contribution due to multiple loci or even the entire genome, or the examination of genetic correlation, or be used as a covariate for the adjustment of genetic effects in a multiple regression model (Wray *et al.*, 2014).

From a statistical perspective, polygenic scores are weighted sums of the genotypes of a set of SNPs.

In most applications of PGS, the weights are usually the SNPs' individual regression coefficients with the phenotype (e.g. Purcell *et al.*, 2009; Wray *et al.*, 2014; Euesden *et al.*, 2015). A critical issue is the total number of SNPs that should be included in the PGS. Although it is usually advisable to use a liberal p-value cutoff in the selection of SNPs to be included, the optimal p-value cutoff is generally unknown (Wray *et al.*, 2014). As a result, in many studies, PGS are constructed using a number of thresholds (Purcell *et al.*, 2009; Ripke *et al.*, 2014; Byrne *et al.*, 2014; Martin *et al.*, 2015; Chang *et al.*, 2015), and there is at least one piece of software developed to facilitate this (Euesden *et al.*, 2015). Generally, we focus on the p-value threshold that achieves the highest correlation/association with the phenotypes in a validation dataset that contains a measure of the phenotype under study. This approach, however, becomes less useful if the phenotype is not available in the target dataset. Recently, Mak *et al.* (2016) sought to overcome this problem by downweighting the usual weights by the SNPs' local true discovery rate, where the additional downweighting or shrinkage factor can be estimated using a data-driven approach. They showed that this leads to comparable predictive performance with the best p-value threshold.

Another issue with this standard approach to PGS calculation is that there is no account taken of the fact that SNPs are in linkage equilibrium (LD) with each other. If SNPs of a particular locus which are in high LD with one another are all included in the score, the contribution to the PGS due to that locus will be exaggerated in the score. For this reason, it is often recommended that SNPs be *pruned* before the application of PG scoring, such that highly correlated SNPs within a locus will have one or more removed (Purcell *et al.*, 2009). Such an approach, however, may well reduce the predictive power of the PGS, as SNPs that are most predictive of the phenotype may be pruned away. A more recent suggestion is that of clumping, which selectively removes less significantly related SNPs to reduce LD (Wray *et al.*, 2014).

In principle, various machine learning methods or Bayesian methods can be applied in the construction of PGS, as they have been applied in the estimation of breeding values in animal studies (Meuwissen *et al.*, 2001; Abraham *et al.*, 2013; Szymczak *et al.*, 2009; Habier *et al.*, 2011; Pirinen *et al.*, 2013; Erbe *et al.*, 2012; Ogutu *et al.*, 2012; Zhou *et al.*, 2013). These methods do not require the assumption of SNP independence or near independence, and have been shown to perform better than simple PGS in

simulation settings. However, their disadvantage is that they cannot be applied to summary statistics. Researchers without access to large datasets are thus unable to take advantage of the power offered by these studies or meta-analyses. A recent development in this direction is Vilhjálmsson *et al.* (2015). The authors proposed an approximate Bayesian method known as LDpred that calculates PGS based on summary statistics, using LD information from a reference panel. Such a development is particularly welcome due to the ready availability of summary statistics from many consortia, often calculated from tens to hundreds of thousands of individuals.

Although Vilhjálmsson *et al.* (2015) demonstrated superior performance of their approach over other simpler methods such as clumping and p-value thresholding through their simulations, a number of issues remain. For example, to what extent is the performance dependent on the choice of the reference panel? The authors recommended the use of a reference panel that is at least 1,000 in sample size, and in their simulations mainly used the validation dataset as the reference panel. Ideally the reference panel from which LD information is derived should share the same ancestry as the data that gave rise to the summary statistics. It is unclear what the consequences are if a reference panel which is not representative of the original population was used instead. Furthermore, the use of Markov Chain Monte Carlo in LDpred means that there is always the possibility of non-convergence in the estimates.

With this in mind, we developed an alternative approach that calculates PGS using summary statistics data, that also takes into account of LD through the use of reference panels. Based on the widely used LASSO and elastic net regression (Tibshirani, 1996; Zou and Hastie, 2005), the method is computationally elegant and fast. In our simulations, we demonstrated that it has good performance even with small reference panels. As such, we can use as reference panels data from such publicly available dataset as the 1000 Genome project, which contains genomic information for many sub-populations across the globe (1000 Genomes Project Consortium, 2015). As with any machine learning approach, a major challenge is in the choice of the tuning parameter. This is particularly difficult when we do not have raw data and hence cannot perform cross-validation. Here, we offer a solution that can potentially be applied more generally. The approach is presented in the methods section and we assessed its performance by simulation studies. Insights gained from the simulations are discussed.

Material and methods

The LASSO problem in terms of summary statistics

Given a linear regression problem

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{X}$ denotes an n-by-p data matrix, and $\boldsymbol{y}$ a vector of observed outcomes, the LASSO is a popular method for deriving estimates of $\boldsymbol{\beta}$ and predictors of (future observations of) $\boldsymbol{y}$, especially in the case where p (the number of predictors/columns in $\boldsymbol{X}$) is large and when it is reasonable to assume that many $\boldsymbol{\beta}$ are zero. LASSO obtains estimates of $\boldsymbol{\beta}$ (weights in the linear combination of $\boldsymbol{X}$) given $\boldsymbol{y}$ and $\boldsymbol{X}$ by minimizing the objective function

$$f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda||\boldsymbol{\beta}||_1 \tag{2}$$

$$= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1 \tag{3}$$

where $||\boldsymbol{\beta}||_1 = \sum_i |\beta_i|$ denote the $\mathcal{L}_1$ norm of $\boldsymbol{\beta}$, for a particular fixed value of λ. In general, depending on λ, a proportion of the β_i are given the estimate of 0. It is also a specific instance of *penalized regression* where the usual least square formulation of the linear regression problem is augmented by a penalty, in this case $2\lambda||\boldsymbol{\beta}||_1$. LASSO lends itself to being used for estimation of $\boldsymbol{\beta}$ in the event where only summary statistics are available, because if $\boldsymbol{X}$ represent standardized genotype data and $\boldsymbol{y}$ standardized phenotype, divided by $\sqrt{n}$, then equation (3) can be written as:

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1 \tag{4}$$

where $\boldsymbol{r} = \boldsymbol{X}^T\boldsymbol{y}$ represents the SNP-wise correlation between the SNPs and the phenotype, and $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$ is the LD matrix, a matrix of correlations between SNPs. As we can obtain estimates of $\boldsymbol{r}$ from summary statistics databases that are publicly available for major diseases/phenotypes (e.g. schizophrenia, depression, bipolar disorders from the Psychiatric Genomics Consortium (`http://www.med.unc.edu/pgc`), height, BMI from the GIANT consortium (`https://www.broadinstitute.org/collaboration/giant/index.p`

and other phenotypes on GWAS Central (`http://help.gwascentral.org/`)), and estimates of LD ($\boldsymbol{R}$) from publicly available genotype such as the 1000 Genome database (1000 Genomes Project Consortium, 2015), equation (4) suggests a method for deriving PGS weights as estimates of $\boldsymbol{\beta}$ by minimizing $f(\boldsymbol{\beta})$.

An issue that surfaces when we substitute $\boldsymbol{R}$ and $\boldsymbol{r}$ with the estimates derived from publicly available data is that the genotype $\boldsymbol{X}$ used to estimate $\boldsymbol{R}$ and $\boldsymbol{r}$ will in general be different. In particular, it will be more appropriate to write $\boldsymbol{R} = \boldsymbol{X}_r^T \boldsymbol{X}_r$ to indicate that the genotype used to derive estimates of LD ($\boldsymbol{X}_r$) will not in general be the same as the genotype that gave rise to the correlations $\boldsymbol{r}$. Writing equation (4) as

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1, \tag{5}$$

however, would imply that (5) is no longer a LASSO problem, because it is no longer a penalized least squares problem. A minimum to (5) can still be sought, although the solutions would often be unstable and non-unique, since $\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ will not generally have a finite minimum.

A natural solution to this problem is to *regularize* equation (5). In particular, if we replace $\boldsymbol{X}_r^T\boldsymbol{X}_r$ with $\boldsymbol{R}_s = (1-s)\boldsymbol{X}_r^T\boldsymbol{X}_r + s\boldsymbol{I}$, for some $0 < s < 1$, then

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R}_s\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1, \tag{6}$$

will be a proper LASSO problem.

Proof. First, we note that $\boldsymbol{R}_s$ is necessarily positive definite for $s > 0$. This means that there always exists $\boldsymbol{X}$ and $\boldsymbol{y}$ such that

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{R}_s, \quad \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{r} \tag{7}$$

Substituting (7) into (6), we see that (6) can be written in a form such as (2) and is therefore a LASSO problem. □

Expanding (6) into

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + (1-s)\boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + 2\lambda||\boldsymbol{\beta}||_1, \tag{8}$$

we note that (8) encompasses a number of submodels as special cases. For example, when $s = 1$,

estimates of $\boldsymbol{\beta}$ will be equivalent to applying soft-thresholding to the univariate correlation summary statistics $\boldsymbol{r}$ (Zou and Hastie, 2005). In particular,

$$\hat{\beta}_i^{s=1} = \text{sign}(r_i)\max(|r_i| - \lambda, 0) \tag{9}$$

This is similar to applying a subset selection to the summary statistics based on p-values, since there is a monotonic relationship between univariate p-values and unsigned correlation coefficients. Another feature is that when $\lambda = 0$, the problem is similar to applying ridge regression to estimate $\boldsymbol{\beta}$, except for a constant scaling value. In most cases, the scale of a PGS is irrelevant, since it is almost never directly used in genomic risk prediction without appropriate scaling (e.g., in So *et al.*, 2011). For a particular choice of s, therefore, equation (8) results in genomic BLUP (Best Linear Unbiased Predictors) (de Los Campos *et al.*, 2013). When $\lambda = 0$ and $s = 1$, the estimated PGS becomes equivalent to simply using the entire set of correlation estimates without shrinkage or subset selection. The flexibility of (8) allows us to examine the effect of modeling LD in estimating weights for PGS through simple simulation studies. In particular, it allows us to examine whether it is always a good idea to take into account of LD information through the use of substitute datasets, as will be demonstrated in our simulations.

Finally, we note that (8) is simply an elastic net problem (Zou and Hastie, 2005), and thus can be solved using fast coordinate descent algorithms (Friedman *et al.*, 2010) for many values of λ at a time. An R package that carries out the estimation of $\boldsymbol{\beta}$ is made available at `https://github.com/tshmak/lassosum`. We made special effort to allow estimation to be done directly on PLINK (Chang *et al.*, 2015) `.bed` files, eliminating the need to load large genotype matrices into R.

Selection of tuning parameters

As with standard LASSO/elastic net problems, in any application, λ and s need to be chosen. In the situation where we have the raw genotype data, these parameters can be selected by cross-validation, although other theoretical approaches are available (Tibshirani, 1996; Efron, 2004; Zou *et al.*, 2007). Without the raw genotype data, however, the task is considerably more difficult. Concerning s, a possible solution is to employ the method of Schäfer and Strimmer (2005), developed for choosing an

optimal shrinkage paramater in the estimation of a covariance matrix. Other more complicated schemes are also possible (Wen and Stephens, 2010), although not pursued in this study.

For the choice of λ, we first note that in the presence of a validation dataset, we can choose λ by maximizing the correlation of the PGS (estimated using different values of λs) with the validation phenotype data, just as it has been done in the choice of a p-value cutoff points in standard PGS calculations (Wray *et al.*, 2014; Euesden *et al.*, 2015). In the absence of a validation dataset, we can simulate this procedure in the following manner, which we refer to as *pseudovalidation* in this paper. First, note that the correlation between a $PGS(\lambda) \equiv \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_\lambda$ and the phenotype $\tilde{\boldsymbol{y}}$ in a new "test" dataset with standardized genotype $\tilde{\boldsymbol{X}}$ is

$$Corr(PGS(\lambda), \tilde{\boldsymbol{y}}) = \frac{\boldsymbol{\beta}_\lambda^T \tilde{\boldsymbol{X}}^T \boldsymbol{P}\tilde{\boldsymbol{y}}}{\sqrt{\boldsymbol{\beta}_\lambda^T \tilde{\boldsymbol{X}}^T \boldsymbol{P}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\lambda \tilde{\boldsymbol{y}}^T \boldsymbol{P}\tilde{\boldsymbol{y}}}} \tag{10}$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n$ is the mean-centering matrix.

In the absence of validation data, $\tilde{\boldsymbol{y}}$ is unavailable. Our solution is to substitute $\hat{\boldsymbol{r}}$ for $\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$, where $\hat{\boldsymbol{r}}$ is a shrunken estimate of the $\boldsymbol{r}$, the observed correlation coefficient vector. Since $\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$ can be interpreted as a correlation coefficient only if $\tilde{\boldsymbol{X}}$ is a standardized genotype matrix and $\tilde{\boldsymbol{y}}$ standardized phenotype, we replace $\tilde{\boldsymbol{X}}$ with its standardized version, $\tilde{\boldsymbol{X}}_0$, and discard the constant $\tilde{\boldsymbol{y}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$ term, so as to maximize the function

$$f(\lambda) = \frac{\boldsymbol{\beta}_\lambda^T \hat{\boldsymbol{r}}}{\sqrt{\boldsymbol{\beta}_\lambda^T \tilde{\boldsymbol{X}}_0^T \tilde{\boldsymbol{X}}_0 \boldsymbol{\beta}_\lambda}} \tag{11}$$

over λ. Here, following Mak *et al.* (2016), we calculated

$$\hat{r}_i = r_i(1 - \mathrm{fdr}_i) \tag{12}$$

where fdr_i is the local false discovery rate of SNP i. While Mak *et al.* (2016) estimated fdr_i using maximum likelihood and a non-parametric kernel density estimator, we found that Strimmer (2008) provided a fast, non-parametric estimator for fdr_i which is constrained to be monotonic decreasing with $|r_i|$, and it is this approach that we have implmented in the simulations.

Some notes on application

In most applications, phenotypes will be binary rather than continuous. Although the theory of this method has been developed with the continuous phenotype in mind, we suggest that when the phenotype is binary, *pseudo-correlation* estimates $\tilde{r}_i$ be derived by converting p-values to correlation, using the monotonic relationship between t-statistics and correlations:

$$\tilde{r}_i = \frac{t_i}{\sqrt{n-1+t_i^2}} \tag{13}$$

In our simulations this resulted in almost identical estimates as using actual (Pearson's product moment) correlations (Figure S1).

Another issue is that in the theory given above, we assume that $\boldsymbol{X}$ and $\boldsymbol{y}$ have been standardized such that $\boldsymbol{r}$ represent the correlation coefficients between the genotype and the phenotype. We note that such standardization can be justified by the fact that the LASSO is often performed on standardized variables (Li *et al.*, 2012; Hastie *et al.*, 2009; Yi *et al.*, 2014). However, when it comes to the construction of PGS, we ought to use unstandardized coefficients as weights. To convert standardized coefficients to unstandardized ones, we can simply use the formula

$$\beta_i^{\text{unstandardized}} = r_i \frac{\text{sd}(\boldsymbol{y})}{\text{sd}(\boldsymbol{X}_i)} \tag{14}$$

where $\text{sd}(\boldsymbol{y})$ and $\text{sd}(\boldsymbol{X}_i)$ are the standard deviations for the phenotype and SNP i.

A third issue concerns the difference between the SNPs with summary statistics and the SNPs that are included in the reference panel. Often the reference panel may not contain all SNPs with summary statistics. Equivalently, there may be no variation within the panel for some SNPs. In LDpred, these SNPs are discarded by default. However, we think that this is not necessary, as it may result in the removal of SNPs that are predictive of the disease/phenotype. An intuitive approach to dealing with these SNPs is that we treat them as if they were all mutually independent and apply soft-thresholding as in (9). Equivalently, we let $\boldsymbol{X}_{ri}$ for these SNPs to be a vector of zero, and we augment equation (8)

by a term $(1-s)\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0$,

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + (1-s)\boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0 + 2\lambda||\boldsymbol{\beta}||_1, \tag{15}$$

where $\boldsymbol{\beta}_0$ denotes the sub-vector of $\boldsymbol{\beta}$ whose $\text{sd}(\boldsymbol{X}_i) = 0$, such that the total ridge penalty for these parameters is 1.

A fourth issue concerns the application of pseudovalidation to clumped data. We proposed above that $\hat{\boldsymbol{r}}$ be estimated using (12) and that the local false discovery rates be estimated using the procedure of Strimmer (2008). An important point is that the method assumes that a sizeable proportion of the $\boldsymbol{r}$ are in fact null. Under clumping, this may not necessarily be the case, and we therefore suggest estimating fdr_i and hence r_i *before* applying clumping.

Simulation studies

We performed a number of simulation studies to assess the performance of our proposed method, which we will refer to as `lassosum`. In our first simulation study, we made use of the Welcome Trust Case Control Consortium (WTCCC) Phase 1 data for seven diseases. We filtered variants and participants using the following QC criteria: genotype rate > 0.99, Minor Allele Frequency > 0.01, Missing genotype per individual < 0.01, SNP rsID included in the 1000 Genome project (Phase 3, release May 2013) genotype data, with matching reference and alternative alleles, on top of the QC done by the original researchers (Wellcome Trust Case Control Consortium, 2007). This resulted in 358,179 SNPs and 15,603 individuals, of which 2,859 were controls. For each of the diseases, there were 1,699-1,902 cases. We randomly chose 1,200 cases and 10,000 controls to form the training set (sampling from all non-cases for that disease rather than just the 2,859 controls), and obtained summary statistics (p-values and signs of log odds ratio) by carrying out SNP-wise logistic regression of the genotype with the phenotype for each of the diseases. We randomly chose 200 cases and 1,000 controls from the remaining individuals as the validation sample. We only used this sample to select the best λ to use in `lassosum`. We assessed the predictive performance of `lassosum` using the Area Under the Curve (AUC) based on another random sample of 200 cases and 1,000 controls from the remaining individuals, which formed the test dataset.

We obtained results using five different reference panels: (a) the validation dataset, (b) the Great Britain (GBR) subsample in the 1000 Genome project ($n = 91$), (c) the European (EUR) subsample in the 1000 Genome Project ($n = 503$), (d) the East Asian (EAS) subsample in the 1000 Genome Project ($n = 503$), and (e) the original training dataset ($n = 11,200$).

In addition to simulations using the seven diseases, we also used a simulated continuous phenotype ($\boldsymbol{y}$) from the standard linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{16}$$

where $\boldsymbol{X}$ is the genotype matrix taken from a sample of 5,600 or 11,200 individuals. In this scenario, 1,000, 5,000, or 25,000 of the 358,179 β_i were given a value randomly sampled from an Exponential distribution, and the heritability of the phenotype ($\hat{\text{Var}}(\boldsymbol{X}\boldsymbol{\beta})/(\text{Var}(\epsilon_i) + \hat{\text{Var}}(\boldsymbol{X}\boldsymbol{\beta}))$ was constrained to be 0.5. As in the previous simulations, we reserved 1,200 individuals for the validation sample and another 1,200 for the test sample. Predictive performance was assessed by the correlation of the PGS with the true predictor, i.e. $Cor(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{X}}\boldsymbol{\beta})$, where $\tilde{\boldsymbol{X}}$ denotes the testing genotype matrix.

We compared our method to LDpred (Vilhjálmsson *et al.*, 2015), a recently developed method which also calculates PGS using summary statistics and a reference panel. LDpred requires a number of parameters as input, including the SNP heritability of the dataset, the size of the "window" around each SNP where LD is calculated, and the a prior proportion of causal SNPs ($P(\text{Causal})$). The first of these can be estimated using LD score regression (Bulik-Sullivan *et al.*, 2015) and this was implemented in the LDpred software. For the second, we followed the recommended practice of using $p/3000$ SNPs on either side of each SNP in calculating the LD information, where p is the total number of SNPs. For the proportion of causal SNPs, we followed Vilhjálmsson *et al.* (2015) in using the following values: 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, and used validation data or our pseudovalidation strategy to select the best proportion of causal SNPs. We note that when $P(\text{Causal}) = 1$, the procedure is equivalent to ridge regression using a tuning parameter, except for the windowing strategy employed in LDpred's computation.

All of the above simulations were repeated four times, except the benchmarking with LDpred, for which we report average prediction performance over 10 repetitions of the simulations.

Because summary statistics are often calculated from large sample sizes and for a large number of

SNPs, we also attempted to carry out simulations using a larger dataset. In particular, we wanted to see whether clumping is an efficient strategy for data reduction, as methods such as LDpred and `lassosum` run quite slowly with a large number of SNPs. However, as we did not have access to raw data of this magnitude, we generated our own simulated data, using the summary statistics available from the GIANT consortium for height as a base (Lango Allen *et al.*, 2010). This dataset contained summary statistics for around 2.5 million SNPs. First, we identified SNPs from the summary statistics that were common with those in the 1000 Genome dataset. We then used Hapgen2 (Su *et al.*, 2011) to generate genotypes, using the CEU (Utah residents with Northern and Western European ancestry) population of the 1000 Genome data as a base. We used two methods for simulating the phenotype. In the first method, 10,000 SNPs were chosen at random among all SNPs, and were assigned effect sizes (β_i) drawn from an exponential distribution. In the second method, we took into account that causal SNPs were likely clustered together and sampled 10,000 causal SNPs in proportion to their local True Discovery Rate (TDR), or 1 minus the local False Discovery Rate, where the local FDR for each SNP was calculated using the method of Strimmer (2008). Because highly significant SNPs were often clustered together (due to LD), in this simulation scheme causal SNPs were clustered. The effect sizes (β_i) of the causal SNPs also followed the exponential distribution. The total heritability of the simulated data was constrained to be 0.45, in agreement with the SNP heritability estimated from Yang *et al.* (2010). Another sample of 1,000 genotypes were simulated in Hapgen2 to form the validation sample, and another 1,000 as the testing sample. We obtained SNP-wise correlation and p-values by linear regression. We applied the *clumping* algorithm of PLINK 1.9 (Chang *et al.*, 2015) to reduce the number of SNPs. Briefly, the clumping algorithm works by first identifying the most signficantly related SNPs, and then deleting SNPs around them that are correlated with them by more than a particular level of r^2 within a particular window. The algorithm finishes when all SNPs are "clumped" into one of these groups represented by a SNP. In our simulations, we set the window size to be 250 kilobases, and the r^2 thresholds to be one of 0.2, 0.5, and 0.8. Because of the time and memory required for this simulation, we were only able to repeat the simulation twice.

In all of the above analyses, we carried out estimation by chromosome.

Results

Although simulations were repeated four times for the WTCCC data and twice for the large simulated data, here we present results from only one of the simulations to simplify presentation and also because results were very similar across repeats. We verified that the observations drawn were consistent across the repeats of the simulations. First, we examined the behaviour of `lassosum` given different shrinkage parameters with respect to the WTCCC data. Figure 1 presents the results for the 7 diseases of the WTCCC in one of our simulations, using the validation dataset for the reference panel. First, we may note that since ridge regression corresponds to the scenario where $\lambda = 0$ in `lassosum`, it is no surprise that the ridge regression prediction performance was very similar to that of `lassosum` when λ was set to 0.001, close to 0. Indeed we may regard the ridge regression results as the asymptotic result of `lassosum` as $\lambda \to 0$. Focusing on the ridge regression results, it is surprising to see that for every disease setting s less than 1 often reduced its predictive power. Ridge regression itself did not appear to be useful in improving PGS prediction, at least for this WTCCC dataset.

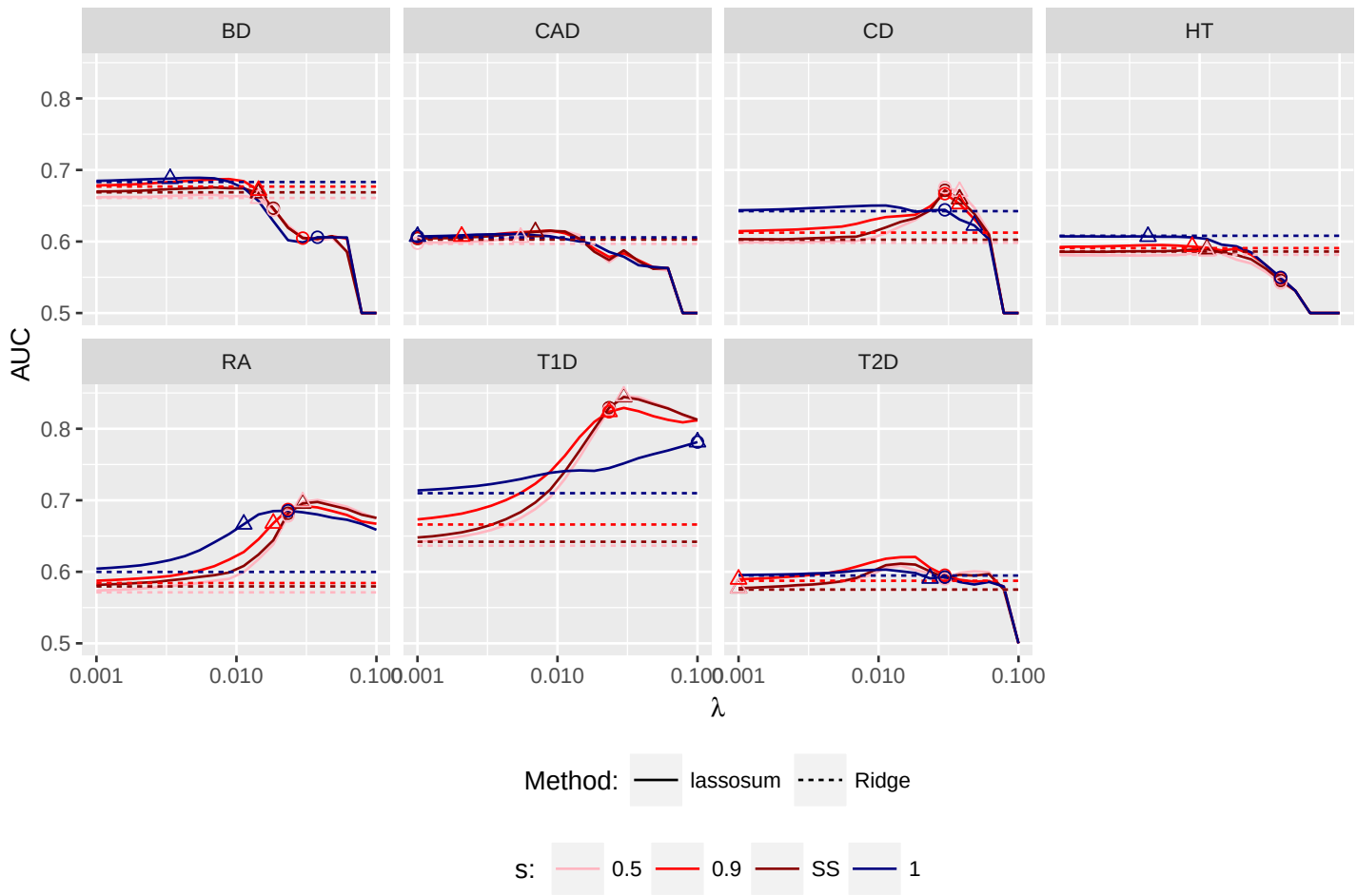

Figure 1: The performance of `lassosum` using WTCCC genotype and phenotypes, using the validation dataset for the reference panel. BD: Bipolar Disorder, CAD: Coronary Artery Disease, CD: Crohn's disease, HT: Hypertension, RA: Rheumatoid Arthritis, T1D: Type 1 diabetes, T2D: Type 2 diabetes, SS: The method of Schäfer and Strimmer (2005). Triangles are the λ value selected using a validation dataset. Circles are values selected using the pseudovalidation strategy proposed in this paper.

However, as λ increased, the usefulness of setting s less than 1 was made manifest for some of the diseases. In particular, for RA and T1D, the best predictive performance was achieved using a λ of around 0.02 to 0.03, and s of 0.5. Using the values of s chosen by the method of Schäfer and Strimmer (2005), which were around 0.44 to 0.80 for these two diseases, also led to similar performance. For the other diseases, however, the value of `lassosum` over simple soft-thresholding (i.e. setting $s = 1$) was not evident. For BD, CAD, T2D, and HT, in particular, the best predictive performance was often achieved by setting $s = 1$ and λ to 0, which is equivalent to the simple PGS strategy using all available SNPs without shrinkage.

An important contribution of this paper is the pseudovalidation strategy, developed for choosing a suitable value of λ in the absence of a suitable phenotype in the validation dataset. It can also be observed that this strategy compared well with using a validation dataset with phenotype and in most cases it was possible to have a value of λ chosen that is close to the optimal. This was especially the case for the diseases RA and T1D, whose signals were strong.

In Figure S2, we present the results using the simulated phenotypes rather than the real phenotypes. Here, we see that when the number of causal SNPs was 1,000, the pattern of the results was similar to what we observed for T1D, i.e., that accounting for LD by setting s to less than 1 improved the prediction of the PGS. When the number of causal SNPs was 25,000, the effect of accounting for LD was less apparent, similar to what we observed for BD, CAD, T2D, and HT. The more causal SNPs there were, the smaller the average effect sizes, and the smaller the signal to noise ratio. Likewise, the larger the sample size, the signal to noise ratio improved, and accounting for LD through setting s to less than one was more useful for the larger sample size of 11,200 than for the smaller sample size of 5,600.

Next, we examined the effect of using different reference panels in prediction. The results using real phenotypes in WTCCC are presented in Figure 2. For all diseases apart from T1D there did not appear to be a consistent pattern in terms of the maximum achievable AUC and the AUC using validation and pseudovalidation. Sometimes using the "wrong" reference panel of EAS (East Asian) led to the best prediction. For T1D however, the results using the EAS was noticeably worse than using the other panels, in terms of the maximum achievable AUC and the AUC attained using validation or pseudovalidation.

This is not surprising given that the WTCCC data was made up of British subjects, and the EAS subset of the 1000 Genome project is from East Asia. On the other hand, the use of the European (EUR) and the Great Britain (GBR) subset both resulted in maximum AUC that is comparable with that obtained using the WTCCC data as reference panel. In general, the use of the original dataset (black line) and the validation dataset (red line) for the reference panel led to very similar predictive performance. Interestingly, the use of the original and the validation dataset as reference panel resulted in a greater difference in predictive power over different values of λ than the use of other reference panels. In particular, for ridge regression, using these reference panels was often worse than the use of other reference panels. Thus it appears the penalty of not using the "correct" λ is higher when using the validation dataset as the reference panel.

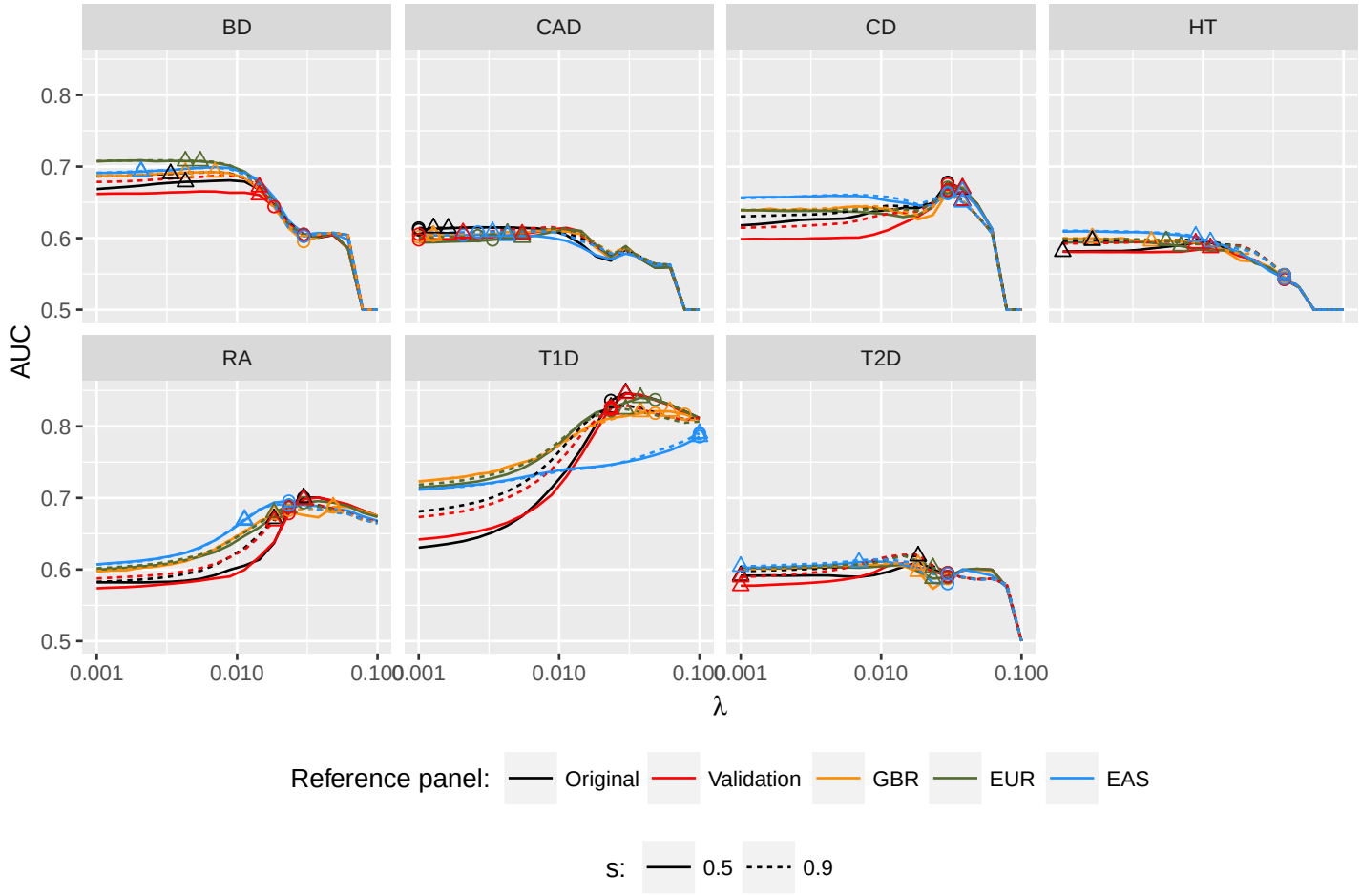

Figure 2: `lassosum` using different reference panels. BD: Bipolar Disorder, CAD: Coronary Artery Disease, CD: Crohn's disease, HT: Hypertension, RA: Rheumatoid Arthritis, T1D: Type 1 diabetes, T2D: Type 2 diabetes. Original: The original dataset that produced the summary statistic; Validation: The validation dataset, which is a sample of 1,200 from the WTCCC data; GBR: The Great Britain subset in the 1000 Genome data; EUR: The European subset in the 1000 Genome data; EAS: The East Asian subset in the 1000 Genome data. Triangles are the λ value selected using a validation dataset. Circles are values selected using the pseudovalidation strategy proposed in this paper.

The results for the same comparisons using simulated phenotypes are presented in Figure S3. Here, the pattern of the results is by and large similar, although the disadvantage of using the EAS subset of the 1000 Genome data as reference panel was less apparent than for T1D above. This was likely because the causal loci were not as clustered together in these simulations, and the choice of reference panel thus had a smaller effect on the predictive power of the PGS.

Next, we benchmarked the performance of `lassosum` to LDpred, a recently proposed PGS method using summary statistics that also accounts for LD (Vilhjálmsson *et al.*, 2015). The results are displayed in Figure S4 for the WTCCC simulations using real phenotypes, and in Figure 3 using simulated phenotypes. For `lassosum`, we used the method of Schäfer and Strimmer (2005) to select the shrinkage parameter. For most of the diseases in the WTCCC dataset, the performance of LDpred, `lassosum`, and simple soft-thresholding (setting $s = 1$ in `lassosum`) was similar, except for T1D, where `lassosum` was superior. For the simulated phenotypes, when the number of causal SNPs was 1,000 and the sample size was 11,200, the performance of LDpred and `lassosum` was similar, and both were superior to soft thresholding. However, when the sample size was halved, the performance of LDpred was drastically reduced. Halving the sample size did not affect `lassosum` in the same way. When the number of causal SNPs was 25,000, and the sample size was 11,200, all methods performed similarly.

Finally we examined the performance of `lassosum` in a large simulated dataset made to have p values distributions that resembled what we observed from the GIANT consortium height GWAS. In Figure S5, we present the qq-plots of the simulated data versus the actual qq-plot generated from the GIANT height summary statistics p-values. When the causal SNPs were randomly drawn across the genome, and when 50,000 samples were simulated, the qq-plot closely resembled that from the GIANT data. Although the GIANT height summary data were derived from around 130,000 participants, we found that we could not increase the sample size further and not drastically decrease the observed p-values. This is likely because the summary statistics from the GIANT consortium were not derived from a simple regression, as our simulated p-values were, but from a genome-wide meta-analysis, and also because genomic inflation control was applied to the summary statistics. When the causal SNPs were sampled in proportion to their true discovery rate, the qq-plot also showed a much greater departure from the null.

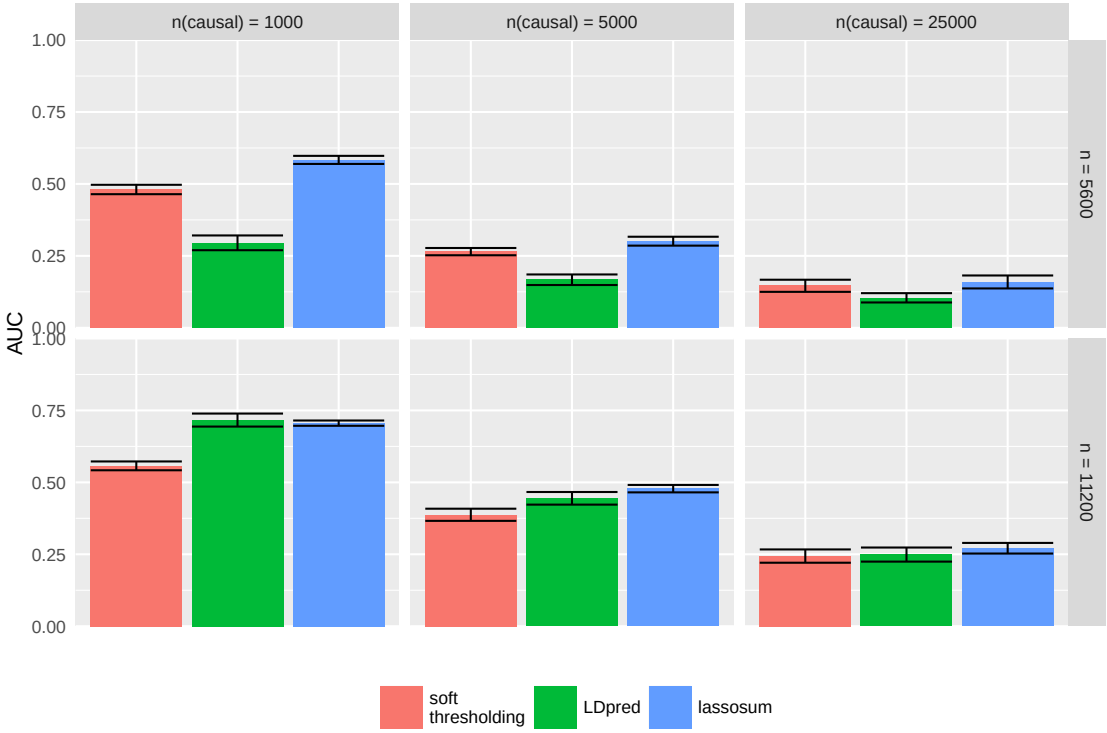

Figure 3: Comparison of `lassosum` with soft thresholding and LDpred using simulated phenotypes. Mean performance averaged over 10 repetitions. The shrinkage parameter of `lassosum` was chosen by the method of Schäfer and Strimmer (2005). Error bars represent 95% confidence intervals.

The main aim of this simulation was to examine the performance of `lassosum` when applied to a large number of SNPs, in this case around 2.5 million, and to see whether pre-filtering by clumping can be an effective method in reducing the number of SNPs in the analysis. In Figure 4, we present the results from this simulation. Although we also assessed the performance of LDpred in this scenario, for reasons unknown to us, LDpred performed rather poorly and we do not show its results here.

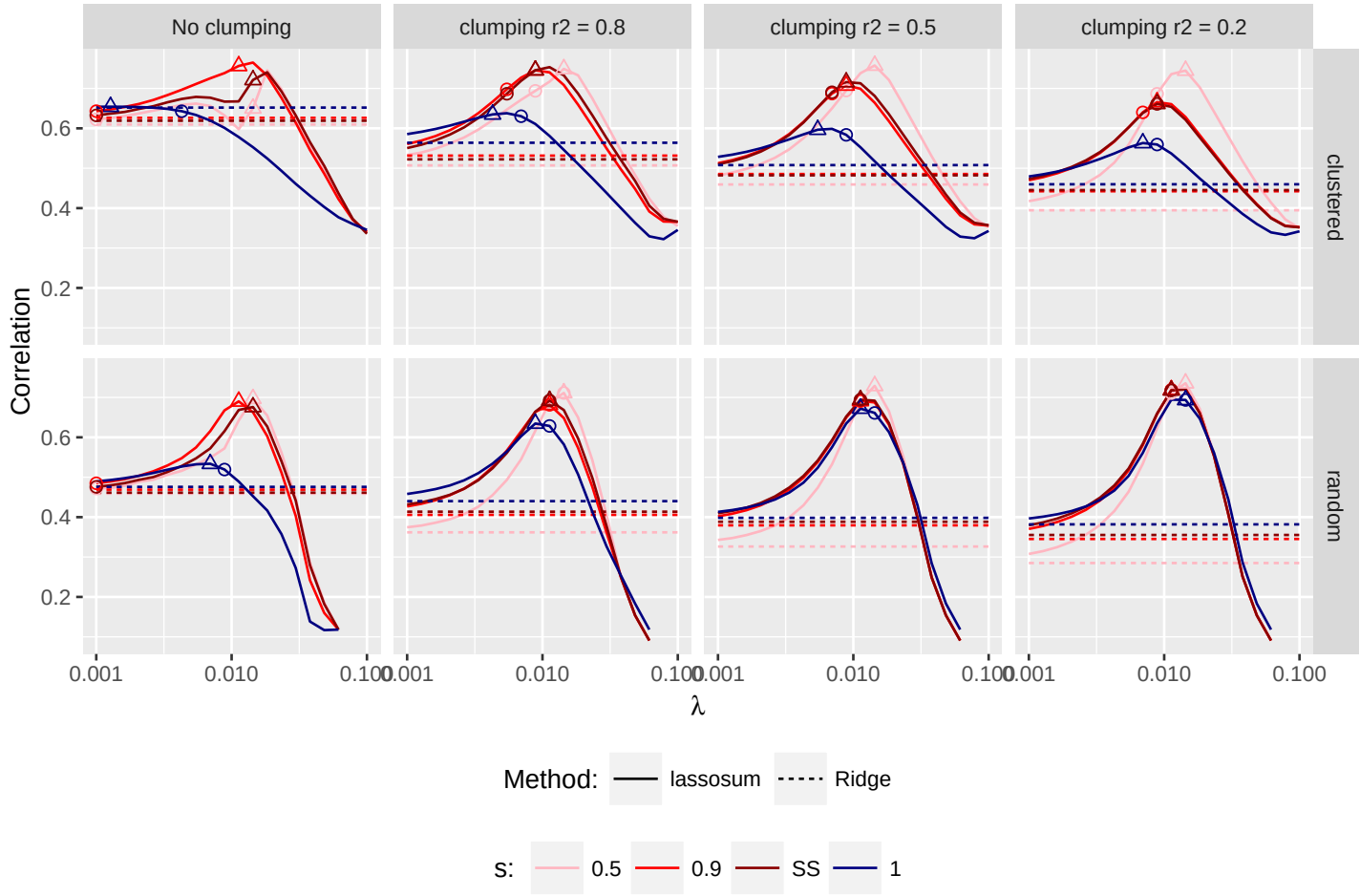

Figure 4: Behaviour of `lassosum` in large simulation dataset as assessed by the correlation with the true predictor in the test dataset. SS: The method of Schäfer and Strimmer (2005). Triangles are the λ value selected using a validation dataset. Circles are values selected using the pseudovalidation strategy proposed in this paper.

First, it can be observed that when the causal SNPs were randomly chosen from the summary statistics loci, and when s was set to 1 (i.e. when LD was not accounted for), clumping was clearly beneficial in increasing the predictive power of the PGS. Interestingly, this benefit remained even when LD was accounted for using `lassosum` by setting s to less than 1, albeit to a lesser extent. When the causal loci were assigned in a clustered manner, and when s was set to 1, clumping in fact reduced the predictive power of the PGS. However, `lassosum` was able to more than compensate for this deficiency. The reduction in predictive power was less when s was set to less than 1. When s was set to 0.5 or set using the method of Schäfer and Strimmer (2005) (SS), there was some irregularity in the performance of the PGS when no clumping was applied. We were uncertain of the reason for this, but the high density of the SNPs and the high levels of correlations observed likely played a role. Moreover, we observed that pseudovalidation failed to select a value of λ close to the best possible in this scenario also when s was less than 1. When s was set to 0.5, the maximum predictive power of the PGS stayed roughly the same across all clumping levels, and was clearly superior to $s = 0.9$ or s set using SS. Nonetheless, pseudovalidation was not as effective in selecting the best possible λ for this level of s.

Discussion

In this paper, we have proposed the calculation of Polygenic Scores using a penalized regression approach using summary statistics and examined its performance in simulation experiments. Our proposed approach, `lassosum`, in general appeared to have a more consistent level of predictive performance than the recently proposed LDpred across a wide variety of simulation settings. This was somewhat surprising to us, as LDpred demonstrated a clear advantage over approaches such as clumping and thresholding in their paper, whereas this was not always clear in our simulations. We think this is due to the smaller sample sizes used in our simulations, both in generating the summary statistics, and in the reference panel. In any case, our results suggest that `lassosum` can be applied using reference panel of size 1,000 or even less, such as those available from the 1000 Genome consortium for specific populations. Indeed the fact that `lassosum` works with reference panels of small sample sizes suggests that it will be useful even when working with raw data, since one can readily speed up LASSO or elastic net estimation by selecting a random subset of the data as reference panel if the number of sample is huge, as our

simulations showed that this does not appear to greatly reduce the predictive power of the resulting PGS.

Moreover, in most cases, the proposed method of pseudovalidation was found to have performance that is almost as good as having validation data for selecting the appropriate λ. We have not examined the issue of selecting the best s to use, although it appears that usually the choice of s has relatively little influence on the performance of the PGS, with s ranging from 0.5 to 0.9 often resulting in similar levels of predictive power. In theory, pseudovalidation can be applied in the selection of s also, although we did not examine this option in details. The method of Schäfer and Strimmer (2005) may be used if an automatic choice is desired, although our simulations using the large simulated dataset suggested that this does not guarantee an optimal choice in terms of predictive power.

Our simulations also suggested that `lassosum` is not necessarily the best method for calculating PGS in all situations. In particular, when the amount of information available in the summary statistics is not great, it appears that often using coefficients from SNPs, without any adjustment for LD, may be the best option. However, even in these scenarios, `lassosum` did not appear to be too far behind than the best option.

Using a reference panel from a population with a totally different ancestry is likely to impair its performance if the SNP signals are strong and correlated. Otherwise, it appears that the choice of reference panel may not drastically affect the performance of `lassosum`.

In the presence of a large number of SNPs, our simulations suggested that clumping can be applied for the pre-filtering of the SNPs. Indeed it seemed necessary if we were to use pseudovalidation to select λ. When causal SNPs are in high LD with one another, clumping may impair the performance of the standard PGS. However, using `lassosum`, this impairment can be avoided or at least reduced. Combining a moderate degree of clumping (e.g. with $r^2 = 0.5$) with a moderate degree of shrinkage (e.g. with $s = 0.5$) in `lassosum` appeared to be a robust choice when dealing with a large number of SNPs.

Some limitations of the present study are worth bearing in mind when considering these results. Real life application of PGS is complicated by the fact that summary statistics may be confounded by population stratification as well as between-population heterogeneity, especially when they are derived

from genome-wide meta-analyses. These complications have not been considered in this paper. In the design of our simulations, we chose to ignore these factors in order that we can understand the behaviour of `lassosum` better. In particular, we have made use of homogenous or nearly homogenous populations in our simulations. One potential problem in using meta-analytic summary statistics is that the original data that generated the summary statistics is an amalgam of datasets across the world, with adjustment for population stratification, and thus there is probably no one single homogenous dataset that is ideal as the reference panel. Further research is needed to clarify what the best strategy is in this situation.

Another important issue concerns the relative merit of estimating PGS using summary statistics data versus using the target dataset alone. When phenotype information is available in the target dataset, conceivably PGS can be applied using the many Bayesian and penalized regression methods that are available (Szymczak *et al.*, 2009; Habier *et al.*, 2011; Zhou *et al.*, 2013; Abraham *et al.*, 2013). Summary statistics from large consortia are supposed to add power to the analyses. However, due to possible between-population differences, summary statistics also contain noise. It is thus not at all certain whether the added information available from the summary statistics can improve the accuracy of the PGS compared to information available from the target population alone.

Finally, in the interest of directing future research in this area, we would like to mention other areas of research that may potentially be merged to improve PGS calculations. Schork *et al.* (2013) showed that different areas of the genomes have different false discovery rate, and therefore different likelihood of being causally associated with a phenotype. Annotation information of the genome can thus potentially be used to improve the predictive power of PGS. Likewise, the fact that many phenotypes have common genetic determinants (pleiotropy) could potentially be exploited to improve PGS. A recent proposal in this direction was given in Li *et al.* (2014). A proposal to combine both annotation information and pleotropy for prioritizing GWAS results is given by Chung *et al.* (2014). There are therefore many potential areas of research in PGS methodology, and we hope that the proposed method in this paper will play a critical role in future developments.

Supplemental Data description

Figure S1 Comparison of PGS constructed using true correlations and pseudocorrelations in the WTCCC dataset

Figure S2 The performance of `lassosum` using WTCCC genotype and simulated phenotypes

Figure S3 `lassosum` using different reference panels with simulated phenotypes

Figure S4 Comparison of `lassosum` with soft thresholding and LDpred

Figure S5 qq-plot for large simulation study

Acknowledgments

We would like to thank Dr. Johnny S. H. Kwan for pointing out to us the work by Strimmer (2008).

Web resources

An R program to implement the methods of this paper is available at `https://github.com/tshmak/lassosum`.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74

Abraham G, Kowalczyk A, Zobel J, and Inouye M (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic epidemiology*, 37(2), 184–95

Agerbo E, Sullivan PF, Vilhjálmsson BJ, Pedersen CB, Mors O, Børglum AD, Hougaard DM, Hollegaard MV, Meier S, Mattheisen M *et al.* (2015). Polygenic Risk Score, Parental Socioeconomic Status, Family History of Psychiatric Disorders, and the Risk for Schizophrenia. *JAMA Psychiatry*, 72(7), 635

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM, Psychiatric Genomics Consortium SWG *et al.* (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3), 291–295

Burgess S and Thompson SG (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4), 1134–1144

Byrne EM, Carrillo-Roa T, Penninx BWJH, Sallis HM, Viktorin A, Chapman B, Henders AK, Pergadia ML, Heath AC, Madden PAF *et al.* (2014). Applying polygenic risk scores to postpartum depression. *Archives of Women's Mental Health*, 17(6), 519–528

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 1–16

Chung D, Yang C, Li C, Gelernter J, and Zhao H (2014). GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLoS Genetics*, 10(11), e1004787

de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, and Sorensen D (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS genetics*, 9(7), e1003608

Domingue BW, Belsky DW, Harris KM, Smolen A, McQueen MB, and Boardman JD (2014). Polygenic risk predicts obesity in both white and black young adults. *PloS one*, 9(7), e101596

Dudbridge F (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3), e1003348

Dudbridge F (2016). Polygenic Epidemiology. *Genetic Epidemiology*, 40(4), 268–272

Efron B (2004). The Estimation of Prediction Error. *Journal of the American Statistical Association*, 99(467), 619–632

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason Ba, and Goddard ME (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science*, 95(7), 4114–29

Euesden J, Lewis CM, and O'Reilly PF (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, (Advanced Access), 1–3

Evans DM, Brion MJA, Paternoster L, Kemp JP, McMahon G, Munaf?? M, Whitfield JB, Medland SE, Montgomery GW, Timpson NJ *et al.* (2013). Mining the Human Phenome Using Allelic Scores That Index Biological Intermediates. *PLoS Genetics*, 9(10)

Evans DM, Visscher PM, and Wray NR (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18(18), 3525–3531

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1–22

Habier D, Fernando RL, Kizilkaya K, and Garrick DJ (2011). Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1), 186

Hastie T, Tibshirani R, and Friedman J (2009). *The elements of statistical learning.* 2nd edition. Springer

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S *et al.* (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317), 832–838

Li C, Yang C, Gelernter J, and Zhao H (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics*, 133(5), 639–650

Li MX, Gui HS, Kwan JSH, Bao SY, and Sham PC (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic acids research*, 40(7), e53

Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, and Kraft P (2011). Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic Epidemiology*, 35(6), 506–514

Mak TSH, Kwan JSH, Campbell DD, and Sham PC (2016). Local True Discovery Rate Weighted Polygenic Scores Using GWAS Summary Data. *Behavior Genetics*, 1–10

Martin J, O'Donovan MC, Thapar A, Langley K, and Williams N (2015). The relationship between common and rare genetic variants in ADHD. *Translational Psychiatry*, 5, e506

Meuwissen TH, Hayes BJ, and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819–29

Ogutu JO, Schulz-Streeck T, and Piepho HP (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC proceedings*, 6 Suppl 2(Suppl 2), S10

Pirinen M, Donnelly P, and Spencer CCA (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1), 369–390

Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, and Posthuma D (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), 702–709

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, and Sklar P (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256), 748–52

Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511, 421–427

Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, Bergen SE, Collins AL, Crowley JJ, Fromer M *et al.* (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10), 1150–9

Schäfer J and Strimmer K (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4, Article32

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Schork NJ *et al.* (2013). All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics*, 9(4), e1003449

So HC, Kwan JSH, Cherny SS, and Sham PC (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *American journal of human genetics*, 88(5), 548–65

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11), 937–948

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FAS *et al.* (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics*, 44(5), 483–9

Strimmer K (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1), 303

Su Z, Marchini J, and Donnelly P (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics (Oxford, England)*, 27(16), 2304–5

Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, and Sun YV (2009). Machine learning in genome-wide association studies. *Genetic epidemiology*, 33(Supplement 1), S51–7

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 58(1), 267–288

Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R *et al.* (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4), 576–592

Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R *et al.* (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics*, 5(10), e1000678

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–78

Wen X and Stephens M (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics*, 4(3), 1158–1182

Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, and Middeldorp CM (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10), 1068–1087

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7), 565–9

Yang J, Manolio Ta, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG *et al.* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics*, 43(6), 519–525

Yi H, Breheny P, Imam N, Liu Y, and Hoeschele I (2014). Penalized Multi-Marker Versus Single-Marker Regression Methods for Genome-Wide Association Studies of Quantitative Traits. *Genetics*, 1–62

Zhou X, Carbonetto P, and Stephens M (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS genetics*, 9(2), e1003264

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320

Zou H, Hastie T, and Tibshirani R (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5), 2173–2192

List of Figures

1 The performance of `lassosum` using WTCCC genotype and phenotypes 15

2 `lassosum` using different reference panels 18

3 Comparison of `lassosum` with soft thresholding and LDpred 20

4 Behaviour of `lassosum` in large simulation dataset 22