Obtaining polygenic transcriptome risk scores (PTRS) directly from GWAS summary statistics

Yanyu Liang[1,*]

1 Section of Genetic Medicine, University of Chicago, Chicago, Illinois, United States of America

* Correspondence to yanyul@uchicago.edu

Abstract

Polygenic Transcriptome Risk Scores (PTRS) are variations of Polygenic Risk Scores (PRS) that use genetically predicted transcriptome as features for prediction instead of directly using genetic variants. We have shown that when PTRS is combined with PRS, they can yield improved prediction performance and portability across populations (Liang et al., 2022). Given the difficulty of training PTRS using large scale individual-level data (due to both computational burden and the lack of data access), we developed a user friendly software that infers PTRS using GWAS summary results and reference LD. We tested three summary statistics-based PTRS approaches: i) Clumping and thresholding (clump-PTRS), keeping trait associated genes while removing highly correlated ones; ii) Summary statistics-based elastic net PTRS (S-EN-PTRS), an extension of lassosum (Mak et al., 2017) to predicted transcriptome; iii) Naive-PTRS, the sum of predicted expressions of significantly associated genes weighted by PrediXcan-estimated effect sizes (Gamazon et al., 2015). Despite reports that individual-level trained elastic net PTRS outperformed clump-PTRS in (Liang et al., 2022), for most of the 11 traits used in the comparison, clump-PTRS outperformed S-EN-PTRS, which outperformed naive-PTRS.

Introduction

PRS is a promising tool to translate GWAS discoveries into clinically-relevant preventive and therapeutic strategies, but there are some challenges that need to be addressed. One issue with PRS is the loss of performance when applied to populations that are underrepresented in current GWASs, i.e. non European-descent groups (Martin et al., 2019). To address this problem, we have proposed PTRS, which uses genetically predicted trancript levels instead of genetic variants as predictors of complex traits (Figure 1B) (Liang et al., 2022). We have shown that PTRS can help improve portability across populations and when combined with traditional PRS, they can yield better performance than PRS alone (Liang et al., 2022).

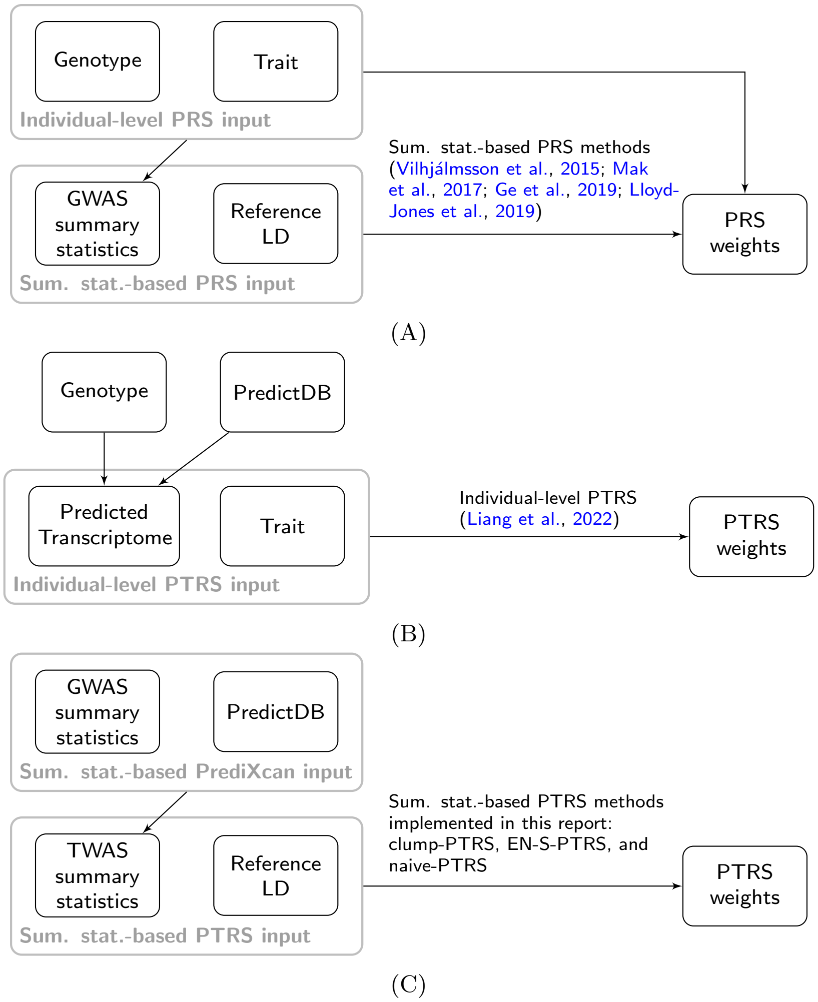
Genotype
Trait
Individual-level PRS input
GWAS summary statistics
Reference LD
Sum. stat.-based PRS input
Sum. stat.-based PRS methods (Vilhjálmsson et al., 2015; Mak et al., 2017; Ge et al., 2019; Lloyd-Jones et al., 2019)
PRS weights
(A)
Genotype
PredictDB
Predicted Transcriptome
Trait
Individual-level PTRS input
Individual-level PTRS (Liang et al., 2022)
PTRS weights
(B)
GWAS summary statistics
PredictDB
Sum. stat.-based PrediXcan input
TWAS summary statistics
Reference LD
Sum. stat.-based PTRS input
Sum. stat.-based PTRS methods implemented in this report: clump-PTRS, EN-S-PTRS, and naive-PTRS
PTRS weights
(C)

Figure 1: **Overview of PRS and PTRS workflows.** **(A)** PRS weights are calculated typically from the GWAS summary statistics and a reference LD (correlation between genetic variants) by summary statistics-based PRS methods (Vilhjálmsson et al., 2015; Mak et al., 2017; Ge et al., 2019; Lloyd-Jones et al., 2019); **(B)** PTRS weights are calculated by (Liang et al., 2022) using individual-level genotype and trait data in the UK Biobank cohort. It first calculates the predicted transcriptome of each individuals from genotype and prediction models of gene expressions (i.e. PredictDB from `predictdb.org`) and then obtains PTRS weights from predicted transcriptome and trait data with a mini-batch elastic net solver tailored to biobank-scale dataset; **(C)** The summary statistics-based PTRS approach implemented in this report uses TWAS summary statistics (obtained from GWAS summary statistics, PredictDB, and a reference LD by summary statistics-based PrediXcan/TWAS (Gusev et al., 2016; Barbeira et al., 2018)) and a reference LD to derive PTRS weights.

Another issue with PRS is that very large training sets (GWASs with very large sample sizes) are needed to obtain relevant prediction accuracy. However, obtaining access to and handling large-scale genetic and phenotypic data can be challenging due to computational and regulatory hurdles. Fortunately, PRS weights can be derived directly from the summary results of GWAS in combination with reference LD information (Figure 1A) (Vilhjálmsson et al., 2015; Mak et al., 2017; Ge et al., 2019; Lloyd-Jones et al., 2019).

In this report, similar to the essence of summary statistics-based PRS approaches, we develop a tool that derives PTRS using GWAS summary statistics and a reference LD (Figure 1C). We implement several approaches to derive PTRS and provide a user-friendly software https://github.com/liangyy/SPrediXcan2PTRS.

Results

To calculate PTRS based on GWAS summary statistics, we extended three common approaches used for PRS construction using GWAS summary statistics: i) clumping and thresholding, ii) summary-based elastic net as implemented in lassosum (Mak et al., 2017), and iii) simple significance thresholding. Clumping and thresholding consists of selecting the most significant features and filtering out features that are highly correlated with each other or do not pass some significance threshold. The lassosum infers the elastic net trained predictors using GWAS summary statistics (Mak et al., 2017). Simple significance thresholding consists of filtering out the variants with non-significant GWAS p-values and using the GWAS effect sizes as the weights in the PRS. We refer to these extensions as clump-PTRS, S-EN-PTRS, and naive-PTRS, respectively.

All three summary statistics-based PTRS approaches start by calculating the TWAS summary statistics using the GWAS summary statistics and transcriptome prediction weights (PredictDB in Figure 1). Clump-PTRS removes genes that are highly correlated and those that do not pass a given p-value threshold. S-EN-PTRS is an implementation of lassosum with TWAS summary statistics taking the place of the GWAS summary statistics. Both the clump-PTRS and S-EN-PTRS require information on the correlation between SNPs which are involved in the prediction of gene expression levels. Naive-PTRS uses the z-scores (or effect sizes) from the TWAS summary statistics as weights for the PTRS. Further details are described in Methods.

To examine the performance of the summary statistics-based PTRS approaches, we selected 11 quantitative trait GWASs listed in Table 1 and calculated the PTRS weights using each of the three approaches described above (Methods). We evaluated the performance of these PTRSs using 5,000 randomly sampled European-descent participants in the UK Biobank. We calculated the Spearman correlation between PTRS scores and the observed UK Biobank traits (Table1) as a measure of the performance of the PTRS (Methods). For the S-EN-PTRS we used a range of shrinkage values (0.01, 0.1, 0.3, 0.5, 0.7, 0.9) and set the mixing parameter $\alpha = 1$ (effectively a lasso penalty was used). For the clump-PTRS, we used a squared correlation cutoff of 0.1 and a sequence of p-value cutoffs (applied after clumping): 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}: 10^{-3}, 0.005, 0.01, 0.05, 0.1, 0.5, 1.

trait short name	ukb field	gwas id	gwas reference	gwas sample size
eosinophil	30150	Astle_et_al_2016_Eosinophil_counts	Astle et al. (2016)	173480
lymphocyte	30120	Astle_et_al_2016_Lymphocyte_counts	Astle et al. (2016)	173480
monocyte	30130	Astle_et_al_2016_Monocyte_count	Astle et al. (2016)	173480
neutrophil	30140	Astle_et_al_2016_Neutrophil_count	Astle et al. (2016)	173480
platelet	30080	Astle_et_al_2016_Platelet_count	Astle et al. (2016)	173480
rbc	30010	Astle_et_al_2016_Red_blood_cell_count	Astle et al. (2016)	173480
wbc	30000	Astle_et_al_2016_White_blood_cell_count	Astle et al. (2016)	173480
bmi	21001	GIANT_2017_BMI_Active_EUR		109433
height	50	GIANT_HEIGHT	Wood et al. (2014)	253288
sbp	4080	ICBP_SystolicPressure	Ehret et al. (2011)	203056
dbp	4079	ICBP_DiastolicPressure	Ehret et al. (2011)	203056

Table 1: **Information on the 11 quantitative traits being used for examining the performance of PTRS.** We used the GWAS harmonized in Barbeira et al. (2021). Column "gwas id" indicates the GWAS ID used in Barbeira et al. (2021). Column "gwas reference" shows the reference to the GWAS. Column "ukb field" shows the Field ID in the UK Biobank

As shown in Figure 2, among most of these 11 quantitative traits, clump-PTRS performed the best, followed by S-EN-PTRS with the performance increasing with higher shrinkage parameters. Naive-PTRS showed the lowest performance.

The lower performance of S-EN-PTRS was unexpected given that the individual-level elastic net-based PTRS outperformed clump-PTRS in our previous study (Liang et al., 2022). One possible explanation of this low performance is the difference in LD between the source GWAS study and the reference one used here. Moreover, we observed that S-EN-PTRS performance increases as the offset term o becomes larger (see definition of the offset in Methods). Since the offset term measures the tendency to treat all genes as independent predictors instead of taking LD information into account, this result also suggested that S-EN-PTRS is potentially sensitive to LD panel.

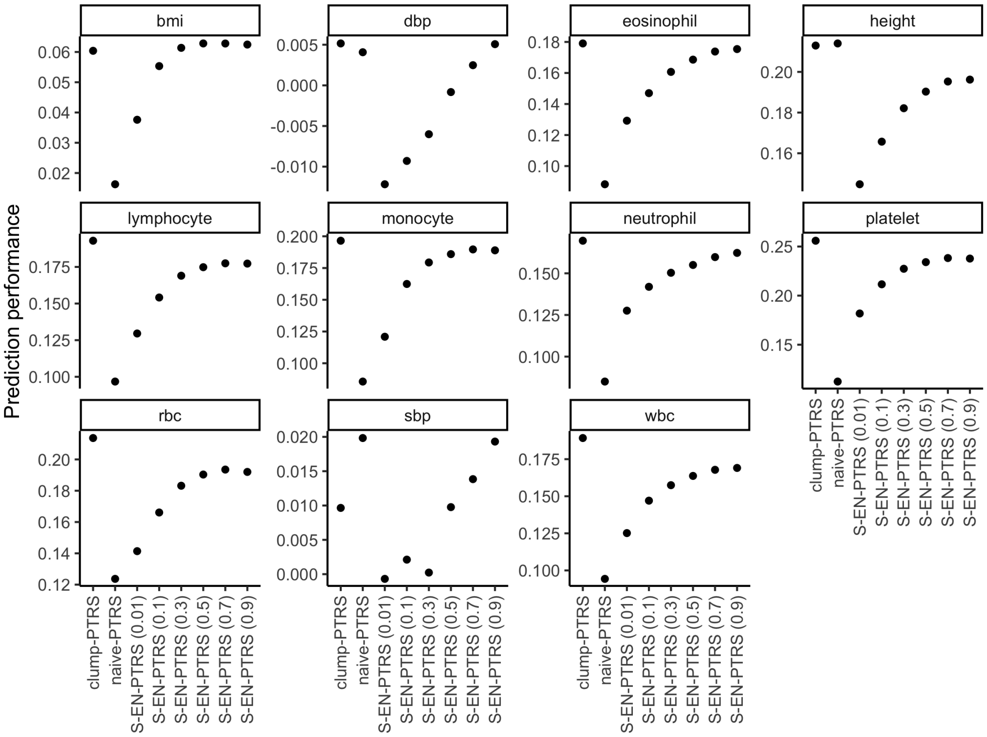

Figure 2: **Performance of S-EN-PTRS, clump-PTRS, and naive-PTRS.** Different summary statistics-based PTRS approaches are shown on x-axis. For S-EN-PTRSs, the value in the parentheses indicates the offset term o being used. PTRSs were evaluated using 5,000 individuals in UK Biobank where half of the samples were used to determine the best-performing hyperparameters (for clump-PTRS, it is the p-value threshold; for S-EN-PTRS, it is the mixing parameter λ). The y-axis shows the Spearman correlation on the other half of the samples with demographic covariates and genetic PCs being adjusted (more details in Methods).

We provide an out-of-box implementation of the three summary statistics-based PTRS approaches, clump-PTRS, S-EN-PTRS, and naive-PTRS, under a unified interface. It takes S-PrediXcan results along with a reference LD panel and generates the corresponding PTRS weights. The software is available at `https://github.com/liangyy/SPrediXcan2PTRS`.

Discussion

In this work, we implemented three summary statistics-based PTRS approaches. We briefly examined the performance of these summary statistics-based PTRSs on 11 quantitative traits while using out-of-sample LD panel. We observed that, with out-of-sample LD panel, clump-PTRS outperforms S-EN-PTRS even though it's been reported that EN-PTRS outperforms clump-PTRS when using individual-level data (Liang et al., 2022). Such discrepancy in performance might be attributed to the LD mismatch since, comparing to clump-PTRS, S-EN-PTRS is more sensitive to the accuracy of the LD panel. To support this claim will require a comprehensive survey of the behavior of these approaches under different scenarios (e.g. variations on the qualities of LD panels, GWASs, and gene expression prediction models). However, a deep performance survey is beyond the scope of this report and we defer it to future work.

Nonetheless, on the basis of the brief performance comparison results in this report, we caution that one should use S-EN-PTRS carefully with an out-of-sample LD panel. With mismatched LD, one could increase the offset parameter in S-EN-PTRS to reduce the effect of mismatch LD on the prediction performance. In practice, one could include both S-EN-PTRS and clump-PTRS as part of their model selection and hyperparameter tuning scheme.

We hope this report along with the software could ease and accelerate the future study and development of PTRS.

Methods

Details of S-EN-PTRS implementation

For N samples and K genes, let $y \in \mathbb{R}^{N \times 1}$ represent the phenotype and let $G \in \mathbb{R}^{N \times K}$ be the predicted expression matrix. As described in (Liang et al., 2022), EN-PTRS is defined by the following problem:

$$\arg\min_x \|y - Gx\|_2^2 + \lambda_0 \cdot (\alpha \|x\|_1 + \frac{1-\alpha}{2}\|x\|_2^2) \tag{1}$$

The EN-PTRS weights correspond to the solution to Eq 1.

We note that the amount of the regularization on each gene depends on the scaling of the corresponding column in G. To resolve such ambiguity, we further assume that each column of G is standardized and so is y. With this assumption, implicitly, the gene-level effect of gene k (x_k) quantifies the amount of phenotypic increase (in the unit of standard deviation of the phenotype) under one unit increase of the predicted expression level of gene k (e.g. the predicted expression goes from the population mean to the population mean plus standard deviation).

Under this scaling, Eq 1 is equivalent to

$$\arg\min_x x'Rx - 2b'x + \lambda \cdot (\alpha \|x\|_1 + \frac{1-\alpha}{2}\|x\|_2^2) \tag{2}$$

R is the sample correlation matrix of G and $b = z_{\text{PrediXcan}}/\sqrt{N}$. Since, equivalently, we use GWAS cohort for EN-PTRS training, N is essentially the GWAS sample size. And $\lambda = \lambda_0/N$.

In practice, we don't have access to R (the in-sample sample correlation). Instead we calculate the sample correlation using a reference LD panel as $\widetilde{R}$. When the GWAS cohort and the reference LD panel are from the same ancestry group, $R \approx \widetilde{R}$ especially when both GWAS and reference LD panel have a lot of samples. In practice, to stabalize the model fitting (similar to lassosum (Mak et al., 2017)), we add an offset term to $\widetilde{R}$ when approximating R. In other words, we let $R = (1-o) \cdot \widetilde{R} + o \cdot I$.

We treat $o \in (0,1)$, $\alpha \in [0,1]$, and $\lambda \in (0,\infty)$ as hyperparameters. For a given pair of o and α, we determine the maximum λ by solving for the minimum possible λ satisfying the KKT condition when $x = 0$ is the solution ((Friedman et al., 2010)). In other words,

$$\lambda_{\max} = \arg\min_{\lambda}\{\lambda : |2R_{k\cdot} \cdot \vec{0} - 2b_k| \leq \alpha\lambda,\ \forall k = 1, \cdots, K\} \tag{3}$$

$$= \frac{2 \cdot \max_k |b_k|}{\alpha} \tag{4}$$

We solve for a sequence of equally spaced (in log-scale) λ values from $\lambda_{\max}$ to $\lambda_{\min}$ and usually $\lambda_{\min} = \lambda_{\max}/100$.

Similar to (Friedman et al., 2010), we solve Eq 2 by coordinate descent. For the sequence $\lambda_1 = \lambda_{\max}, \lambda_2, \cdots, \lambda_J = \lambda_{\min}$, we first set $x = \vec{0}$ for $\lambda = \lambda_{\max}$. And then, for each $\lambda = \lambda_j$, we initialize x as the solution from $\lambda = \lambda_{j-1}$.

Additionally, to obtain $\widetilde{R}$, we need information more than the one required for S-PrediXcan, since we need to know both the variant/variant covariance for each variant pair within a gene and the variant pair from two different genes. We note that it is not suitable to pre-compute gene covariance since, usually, the GWAS may miss some genetic variants in the prediction models and we should exclude these variants for the PTRS training. Instead, similar to the S-PrediXcan implementation, we calculate $\widetilde{R}$ on the fly using the pre-computed variant/variant covariance from the reference LD panel. In practice, we consider gene/gene covariance for each chromosome and assume the covariance between genes from different chromosomes equals to zero. With this assumption, we need to pre-compute and store variant/variant covariance for all the variants that appear in at least one of the prediction models and lie in the same chromosome. For chromosomes 1 to 22, the number of variants to consider ranges from 4,000 to 30,000, for which we need to store all the pairwise information. In fact, the number of samples in the reference LD panel N_{LD} is usually smaller than the number of variants being considered N_{SNP} and it is an order of magnitude smaller when using GTEx data as the reference (sample size is about hundreds) The N_{SNP}-by-N_{SNP} variant/variance covariance matrix can be represented by only N_{LD} eigenvalues and eigenvectors. We make use of this property and save the eigenvalue decomposition of the variant/variant covariance matrix instead.

In principle, we need to loop over the genome-wide variants until convergence. But here, we also implement an alternative scheme in which we loop over variants within each chromosome (this is similar to lassosum (Mak et al., 2017) which solves elastic net models for each LD block while using the same y for all blocks). We note that this chromosome-wise approach usually gives similar results as the genome-wide iteration and converges a little faster.

Details of clump-PTRS implementation

Given a list of genes with PrediXcan z-scores and r^2, a cutoff on squared correlation between genes, we want to perform clumping such that genes with duplicated signals (genes which are "highly" correlated with a more significant gene) are removed. The procedure is similar to (Liang et al., 2022) and it is outlined in Algorithm 1.

```
Algorithm 1: Clumping procedure
Input  : Z-scores for a list of predictor candidates $z \in \mathbb{R}^{K \times 1}$, correlation between predictors
         $R \in \mathbb{R}^{K \times K}$, a squared correlation cutoff $r^2$.
Output: A list of predictors (in terms of index) among which $R_{ij}^2 < r^2$ and the ones with high $|z|$
         should always be preferred.
Function Clumping($z, R, r^2$):
    sortedIndex = argsort($|z|$, decreasing);
    Initialize a Hash table statusHash with keys $1, \cdots,$ K and value = 0;
    Initialize outputList as empty;
    for $j \leftarrow 1$ to $K$ do
        $i$ = sortedIndex[$j$];
        if statusHash[$i$] = -1 then
            continue;
        Append $i$ to outputList;
        for $k \leftarrow j$ to $K$ do
            if statusHash[$k$] = -1 then
                continue;
            end
            if $R_{ik}^2 \geq r^2$ then
                statusHash[$k$] = -1;
            end
        end
    end
    return outputList
End
```

Applying summary statitics-based PTRS to 11 GWASs

We applied clump-PTRS, S-EN-PTRS, and naive-PTRS to 11 GWASs listed in Table 1. These GWASs were harmonized previously and imputed in GTEx v8 WGS samples (Barbeira et al., 2021). We used the CTIMP models and cross-tissue gene expression imputation models (Hu et al., 2019) from Whole Blood, which were trained on GTEx v8 European samples (Barbeira et al., 2020). The variant/variant covariance was obtained from GTEx v8 European samples.

Evaluating PTRS performance using UK Biobank data

We evaluated the prediction performance by calculating the Spearman correlation between the predicted and observed phenotype values on 5,000 randomly selected UK Biobank participants with European ancestry. To take covariates into account, we regressed out the top 10 genetic PCs, age, and sex (and second order terms of age and sex) from both the observed and the predicted values before calculating the correlation. The corresponding UK Biobank phenotypes used for the evaluation are shown in Table 1.

To select the hyperparameters, we split the selected participants into two halves. We used the first half samples to determine the hyperparameters (λ for S-EN-PTRS and the p-value cutoff for clump-PTRS) by selecting the best-performing models. And then we used the second half of the samples to calculate the prediction performance under the models with the selected hyperparameters.

Code Availability

The software implemented for this work is available at https://github.com/liangyy/SPrediXcan2PTRS.

Acknowledgment

We thank Hae Kyung Im for helpful discussion. We thank Natasha Santhanam and Beau Burnett for helping us edit the report. This research has been conducted using the UK Biobank Resource under Application Number 19526.

References

W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429, 2016.

A. N. Barbeira, S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1–20, 2018.

A. N. Barbeira, O. J. Melia, Y. Liang, R. Bonazzola, G. Wang, H. E. Wheeler, F. Aguet, K. G. Ardlie,

X. Wen, and H. K. Im. Fine-mapping and qtl tissue-sharing information improves the reliability of causal gene identification. *Genetic Epidemiology*, 2020.

A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, et al. Exploiting the gtex resources to decipher the mechanisms at gwas loci. *Genome biology*, 22(1):1–24, 2021.

G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S.-J. Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

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

E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

T. Ge, C.-Y. Chen, Y. Ni, Y.-C. A. Feng, and J. W. Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1–10, 2019.

A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. De Geus, D. I. Boomsma, F. A. Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

Y. Hu, M. Li, Q. Lu, H. Weng, J. Wang, S. M. Zekavat, Z. Yu, B. Li, J. Gu, S. Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics*, 51(3):568–576, 2019.

Y. Liang, M. Pividori, A. Manichaikul, A. A. Palmer, N. J. Cox, H. E. Wheeler, and H. K. Im. Polygenic transcriptome risk scores (ptrs) can improve portability of polygenic risk scores across ancestries. *Genome biology*, 23(1):1–18, 2022.

L. R. Lloyd-Jones, J. Zeng, J. Sidorenko, L. Yengo, G. Moser, K. E. Kemper, H. Wang, Z. Zheng, R. Magi, T. Esko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11, 2019.

T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480, 2017.

A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584, 2019.

B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P.-R. Loh, G. Bhatia, R. Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592, 2015.

A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, Z. Kutalik, N. Amin, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.