Assumptions about frequency-dependent architectures of complex traits bias measures of functional enrichment

Shadi Zabad[1], Aaron P. Ragsdale[2], Rosie Sun[2], Yue Li[*,1,3], and Simon Gravel[**,2,3]

[1]Department of Computer Science, McGill University, Montreal, QC, Canada
[2]Department of Human Genetics, McGill University, Montreal, QC, Canada
[3]Quantitative Life Science, McGill University, Montreal, QC, Canada

Abstract

Linkage-Disequilibrium Score Regression (LDSC) is a popular framework for analyzing GWAS summary statistics that allows for estimating SNP heritability, confounding, and functional enrichment of genetic variants with different annotations. Recent work has highlighted the influence of implicit and explicit assumptions of the model on the biological interpretation of the results. In this work, we explored a formulation of LDSC that replaces the r^2 measure of LD with a recently-proposed unbiased estimator of the D^2 statistic. In addition to modest statistical difference across estimators, this derivation highlighted implicit and unrealistic assumptions about the relationship between allele frequency, effect size, and annotation status. We carry out a systematic comparison of alternative LDSC formulations by applying them to summary statistics from 47 GWAS traits. Our results show that commonly used models likely underestimate functional enrichment. These results highlight the importance of calibrating the LDSC model to achieve a more robust understanding of polygenic traits.

[*]Corresponding author. E-mail: yueli@cs.mcgill.ca
[**]Corresponding author. E-mail: simon.gravel@mcgill.ca

1 Introduction

Linkage-Disequilibrium Score Regression (LDSC) provides a general framework for understanding the architecture of polygenic traits from GWAS summary statistics [4, 12]. Because of its tractability and computational efficiency, it has become a central tool to estimate statistical and population genetic quantities, such as SNP heritability (h^2_{SNP}) and confounding in GWAS results [4], functional enrichment [5], cross-trait genetic correlations [3] as well as genetic correlations across ethnic groups for the same trait[16]. This diversity of applications stems from the ease of incorporating various assumptions about the architecture of a trait in the LDSC framework. In its original formulation, the LDSC model made the implicit assumption that the expected phenotypic variance explained by each genetic variant was independent of allele frequency. However, subsequent work has shown that this assumption is restrictive and results in downwardly biased estimates of heritability [18].

To guard against these potential sources of bias, the trend in the field has been towards incorporating more expressive models that describe the expected squared effect of a variant as a function of different genetic annotations [5, 6, 7, 10, 11], such as MAF decile bins and whether a SNP occurs in an annotated enhancer region. An advantage of this approach, commonly referred to as Stratified LDSC (S-LDSC), is that it allows for arbitrary relationships between allele frequency and mean squared effect sizes, thus reducing the need for a priori assumptions that can bias results. It also provides measures of functional enrichment for the categories incorporated into the model. Despite these many advantages, previous work has shown that the way we model the dependence of effect size on functional annotations can result in substantially different estimates of partitioned heritability and functional enrichment [8, 18, 19].

In this work, we reformulate the LD Score regression framework by replacing the r^2 measure of LD with an estimate based on the D^2 statistic, which has desirable statistical properties [14]. Even though the choice of the LD statistic itself has little effect on heritability estimates, this formulation highlights implicit assumptions of commonly-used stratified models. Specifically, these models imply relationships between allele frequency, effect size, and annotation status that do not agree with empirical observations. While these assumptions have a small impact on global estimates of SNP heritability, we show that they can result in systematic biases in estimates of functional enrichment. To correct for these biases, we propose using modified S-LDSC models that better capture the empirical relationship between allele frequency and effect size for annotated variants.

2 Materials and Methods

2.1 LD Score Regression as a function of D^2

In our formulation of LD Score Regression, we assume a standard linear polygenic model $Y = X\beta + \epsilon$, where Y is a vector of phenotypes for N individuals, X is a $N \times M$ mean-centered (but not variance-normalized) matrix that encodes the genotype at M SNPs, β is a vector of random and independent effect sizes, and ϵ is a vector of random and independent environmental and other effects. Under this model, assuming no confounding, the expectation of the χ^2 association statistic at locus j can be expressed as a function of the squared covariance in allele frequency D^2_{jk} at variants j and k as:

$$\mathbb{E}[\chi^2_j] \approx N \sum_k \frac{4}{\mathrm{Var}(X_j)} D^2_{jk}\, \mathrm{Var}(\beta_k) + 1. \tag{1}$$

Here, $\mathrm{Var}(\beta_k)$ is the variance of causal effect size at SNP k and $Var(X_j)$ is the variance in genotype across individuals at the focal SNP j. In LDSC, we seek to build a parameterized model for $\mathrm{Var}(\beta_k)$ that can be fitted to the observed χ^2. Since $\mathrm{Var}(\beta_k)$ is the mean squared effect size for variant k, we can then sum over all SNPs to obtain an estimate of SNP heritability. While the simplest model takes $\mathrm{Var}(\beta_k)$ to be a constant across SNPs, various models have been proposed that take into account allele frequencies as well as functional annotations. Here we consider models of the form

$$\begin{aligned} &\mathrm{Var}(\beta_k) = \mathrm{Var}(X_k)^{-\alpha} \left[\tau_0 + \theta \left(\sum_{c=1}^{96} \tau_c a_c(k) \right) \right] \\ &\alpha \in \{0, 0.25, 0.5, 0.75, 1\}, \quad \theta \in \{0, 1\}. \end{aligned} \tag{2}$$

The mean squared effect size $\mathrm{Var}(\beta_k)$ of SNP k is defined in terms of two components: a dependence on minor allele frequency (MAF) through the parameter α, and a dependence on a set of 96 functional annotations $a_c(k)$ that were incorporated into the baseline-LD model (v2.2) ([6, 8], see Web Resources). This dependence is controlled by the parameter θ. If $\theta = 0$, we recover the univariate models introduced in Bulik-Sullivan et al. [4]. If $\theta = 1$ we obtain the stratified models that are most commonly used in more recent work [5, 6, 7, 11, 10].

The contribution of each annotation a_c to SNP heritability is modulated by the corresponding parameter τ_c, which is estimated in the LDSC regression procedure. In this formulation, τ_0 corresponds to the "base" annotation that includes all SNPs (e.g. $a_0(k) = 1$ for all k). Table 1 summarizes some of the relevant models that will be analyzed in this paper.

In previous work, α has often been treated as a fixed parameter, usually set to $\alpha = 1$. In the univariate case ($\theta = 0$), this is a mathematically convenient choice that implies constant variance explained by all SNPs and leads to simple expressions of LD scores in terms of the LD statistic r^2 (as detailed below). In

Heritability model	**Parameters**	**Comments**
GCTA model [22]	$\alpha = 0; \theta = 0$	The expected mean squared effect size is a constant.
Constant-variance model [4]	$\alpha = 1; \theta = 0$	The expected variance explained by each SNP is constant.
Baseline-LD model (v2.2) [8]	$\alpha = 1; \theta = 1$	A stratified LDSC model that includes a total of 96 functional annotations, including continuous LD-related annotations and MAF bins. This model gives more weight to annotations found in rare variants.
MAF-independent baseline-LD model	$\alpha = 0; \theta = 1$	Similar to the baseline-LD model, though the effect of each annotation is independent of allele frequency.

Table 1: The SNP-heritability models analyzed in this paper.

turn, this implies that rare variants have much larger per-allele biological effect than common variants. In other research settings, α is treated as a continuous parameter that can be fit to data to measure the strength of negative selection for a given trait. Average inferred values of α for a number of UK Biobank traits ranged from 0.2 to 0.5, with a mean value of roughly 0.38 [25, 15, 18]. However, these values of α were inferred for the univariate setting and it is not clear if the interaction between MAF and the various functional annotations can still be characterized by these global estimates. In fact, a recent report using an updated version of the LDAK and SumHer models [19, 17, 20] found variations in the α values ranging from 0.23 to 0.67 for SNPs in all but one functional category, and 0.25 for all variants.

Despite these observations, in standard applications of the stratified LDSC models ($\theta = 1$), it is still commonly assumed that $\alpha = 1$ [6, 7, 10, 11]. The choice $\alpha = 1$ is not particularly biologically plausible — it is a leftover from the mathematically convenient formulation of LDSC in terms of the r^2 statistic. Equation (2) shows that the assumption $\alpha = 1$ in the stratified setting implies that the contribution of each annotation to predicted effect size is much larger for rare variants. Put another way, under the $\alpha = 1$ model, rare variants in a given functionally enriched category c are predicted to have much larger per-allele biological effect than common variants in the same category. Since this assumption has been found to not be the valid in genome-wide analyses [25, 15], there is no compelling reason to think that it holds within specific categories of variants.

3 Results

3.1 Comparing LD scores across models and estimators

The family of models outlined in Equations (1) and (2) may be equivalently expressed in terms of LD scores:

$$\mathbb{E}[\chi_j^2] \approx N\left[\tau_0 \ell(j,0) + \theta \sum_{c=1}^{96} \tau_c \ell(j,c)\right] + 1, \tag{3}$$

where the LD score $\ell(j,c)$ of variant j and category c is defined as the sum of the annotation value $a_c(k)$ multiplied by a MAF-weighted linkage disequilibrium term $L_{jk}(\alpha)$ over all neighbouring SNPs k:

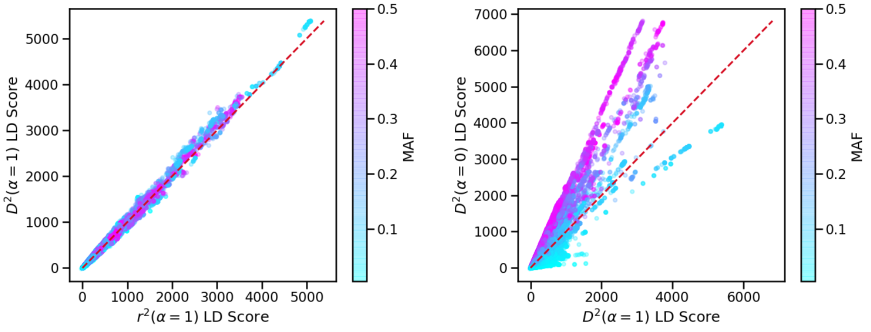
$$\begin{aligned} \ell(j,c) &:= \sum_k L_{jk}(\alpha) a_c(k) \\ L_{jk}(\alpha) &= \frac{4D_{jk}^2}{\operatorname{Var}(X_j)} \operatorname{Var}(X_k)^{-\alpha} = r_{jk}^2 \operatorname{Var}(X_k)^{1-\alpha}. \end{aligned} \tag{4}$$

In standard applications of LDSC, the MAF-weighted LD measure $L_{jk}(\alpha)$ is typically taken to be r_{jk}^2, which implies $\alpha = 1$. However, for arbitrary α, it can be expressed more generally in terms of a product of the variances at SNPs j and k with the squared covariance D_{jk}^2 or the squared correlation r_{jk}^2. This generalized measure can be estimated in-sample or from a reference panel using, e.g, the following two estimators:

$$\begin{aligned} \hat{L}_{jk,D} &= \frac{4\widehat{D_{jk}^2}}{\widehat{\operatorname{Var}(X_j)}} \widehat{\operatorname{Var}(X_k)}^{-\alpha} \\ \hat{L}_{jk,r} &= \widehat{r_{jk}^2} \widehat{\operatorname{Var}(X_k)}^{1-\alpha} \end{aligned}$$

where $\widehat{D_{jk}^2}$ is the unbiased estimator of [14] and $\widehat{r_{jk}^2}$ is the bias-corrected estimator used in the original LDSC model [4, 24]. The estimator for $\hat{L}_{jk,r}$ has a particularly simple form for $\alpha = 1$, which is probably a reason for the popularity of this parameter choice. To evaluate whether choices of estimators affect LDSC regression results, we analyzed the differences between the two statistical estimators across different values of α. To perform this analysis, we computed LD scores according to Equation (4) using both estimators and α values of $\{0, 0.25, 0.5, 0.75, 1\}$ for three of the super populations in the 1000 Genomes Project (Africans, Asians and Europeans) [1, 21].

Both estimators produce highly correlated estimates ($R^2 = 0.99$) (Figure 1 (a)), and their bias is comparable even for small panel sample size (Supplementary Material, Supplementary Figure 1). The consistency of these two estimators suggests that they perform equally well in estimating LD scores irrespective of the value of α. Thus, biological realism might be a more relevant



(a) r^2 vs. D^2

(b) $\alpha = 1$ vs. $\alpha = 0$

Figure 1: Genome-wide comparison of LD Scores in the European samples ($N = 489$) in the 1000 Genomes Project. **(a)** shows the distribution of LD Scores obtained from the r^2 (x-axis) and D^2 (y-axis) estimators. In **(b)** we show the normalized LD Scores from the D^2 estimator but computed according to different models of SNP heritability ($\alpha = 1$ on the x-axis and the $\alpha = 0$ on the y-axis). Points are colored by the focal SNP's minor allele frequency (MAF).

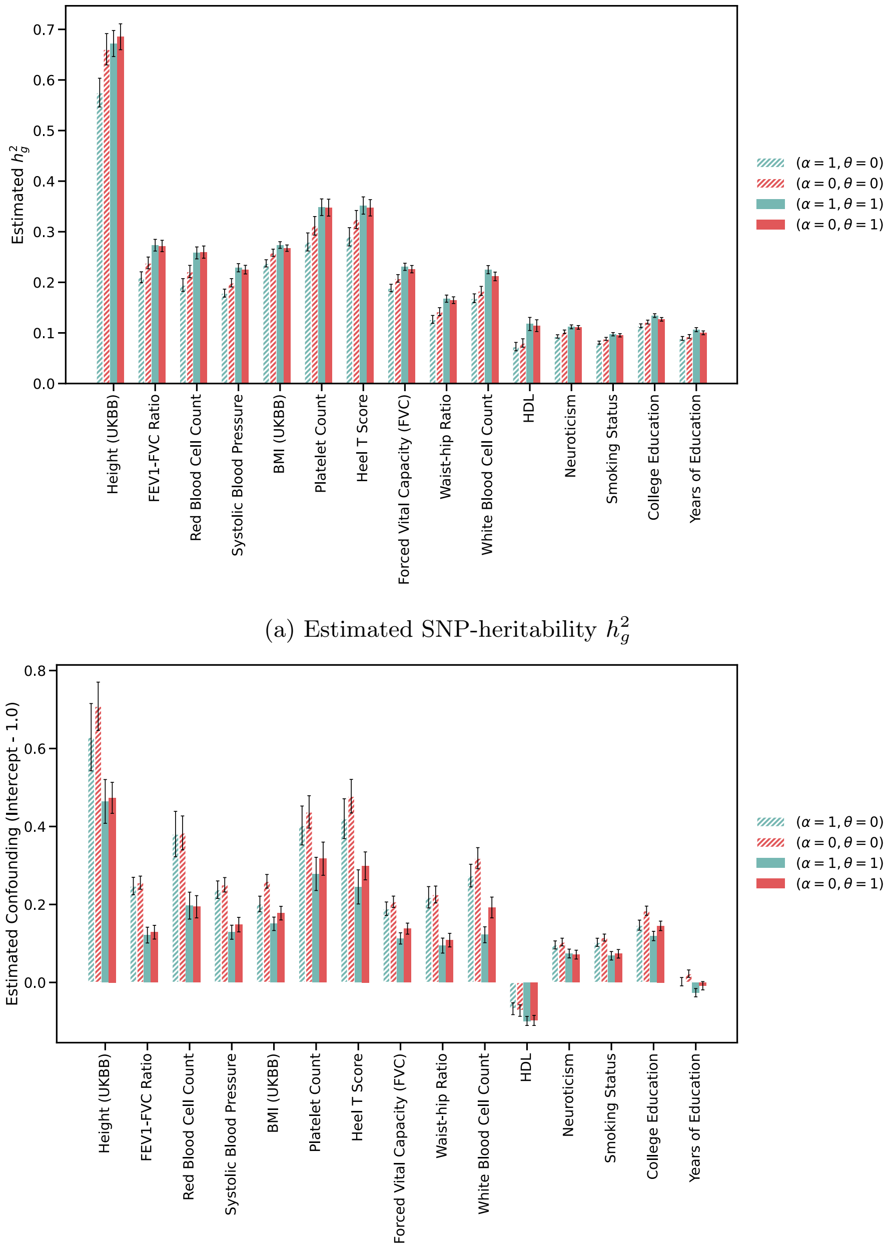
criterion when choosing α than statistical convenience. For the remainder of the discussion in the main text, we will use the $\hat{L}_{jk,D}$ estimator.

Next, we sought to highlight the influence of model assumptions on the LD scores computed from the 1000 Genomes data. For clarity of exposition, we focus on the univariate case where the LDSC model can be written as: $\mathbb{E}[\chi_j^2] \approx N\tau_0\ell(j, 0) + 1$ and the LD score is simply the sum of the generalized LD measures $L_{jk}(\alpha)$. To qualitatively compare the LD scores with different values of α in a consistent manner, we normalize them such that the slope of the univariate regression becomes $\frac{N}{M}h_{SNP}^2$, where h_{SNP}^2 is the total heritability (see Appendix C.1).

As expected, Figure 1(b) shows that common variants have higher LD scores under the $\alpha = 0$ model compared to the $\alpha = 1$ model, and the opposite trend is observed for rare variants. This effect is less stark for intermediate values of α (Supplementary Figure 2). In practical terms, the choice of α determines which categories of SNPs will have higher weight in the LDSC regression and, as has been documented before [18], this will in turn influence the global estimates of SNP heritability.

3.2 Estimates of SNP-heritability and Confounding for 47 GWAS Traits

To understand the implications of choosing different values of α on global estimates of SNP heritability and confounding, we applied the LD Score regression framework to a total of 47 GWAS traits (41 independent traits), for which summary statistics have been previously analyzed [8, 11] (Supplementary Table S1,



(a) Estimated SNP-heritability h_g^2

(b) Estimated confounding (Intercept - 1)

Figure 2: Estimates of SNP heritability and confounding for 15 GWAS traits using 4 different models of heritability. Dashed bars show estimates for univariate models. Solid bars show estimates for stratified models. Error bars correspond to jackknife standard errors. Color code: turquoise ($\alpha = 1$), red ($\alpha = 0$).

see Web Resources). We confirmed that the choice of statistical estimator has a modest impact on estimates of global heritability and confounding (Supplementary Table S7): for example, the $\hat{L}_{jk,D}$ estimator in the $\alpha = 1$ stratified model gave lower estimates of heritability than the $\hat{L}_{jk,r}$ estimator for all 47 traits, but the relative differences were always smaller than 6% (Supplementary Table S3), and always smaller than twice the block-jackknife standard error of the estimator. By contrast, the choice of α can produce large differences.

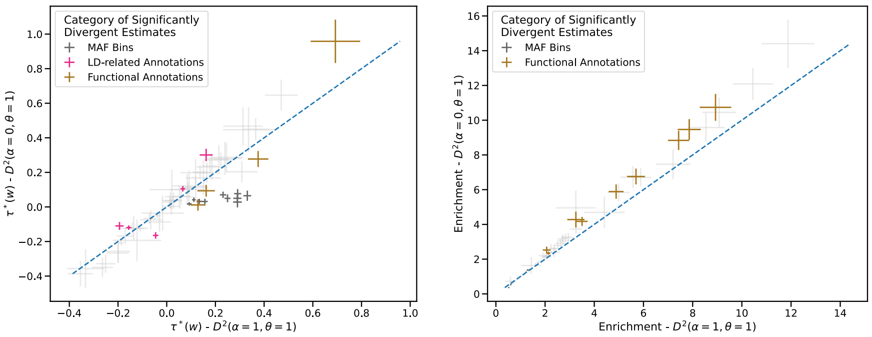
For the univariate models ($\theta = 0$), we find that estimates of heritability with $\alpha = 0$ are different by more than two standard errors to the $\alpha = 1$ estimate for 29 out of the 47 traits considered (Figure 2, Supplementary Figure 6, Supplementary Tables S5-6). In general, we observe that the univariate models with $\alpha = 0$ produce heritability estimates that are closer to the estimates from the stratified models for 44 out of 47 traits (Figure 2(a), Supplementary Tables S2-4), which is consistent with previous estimates of α values that are closer to zero ($\alpha \approx 0.38$, [25, 15]). This trend is reversed for the intercept, with the $\alpha = 1$ model producing estimates that are closer to the stratified models for most traits. Supplementary Figure 6(b) shows that intermediate values of α, such as $\alpha = 0.25$, produce estimates of confounding and global heritability that are closer to the stratified models than both $\alpha = 0$ and $\alpha = 1$.

In the case of the stratified models ($\theta = 1$), the differences are more subtle. Out of the 47 traits analyzed, none showed a significant difference in the global estimates of SNP heritability across the models with $\alpha = 0$ and $\alpha = 1$ (Supplementary Tables S2-4). At the same time, small but significant differences are observed in the estimates of the intercept for 3 of the traits analyzed (Eosinophil Count, Tanning, and White Blood Cell Count) (Supplementary Tables S2-4). Overall, our analysis confirms that the stratified models (with the MAF decile bins included) successfully counteract the bias induced by an arbitrary choice of α, producing global estimates of SNP heritability and confounding that are largely concordant across different choices of α (Figure 2, Supplementary Figure 7).

3.3 Examining Estimates of Coefficients and Functional Enrichment

The preceding analyses showed that the stratified models generally produce concordant estimates of global parameters such as SNP heritability and confounding, with the choice of α having only a minor impact. However, as has been noted above, different choices of α can still result in different estimates for quantities associated with partitioned heritability, such as standardized heritability coefficients and functional enrichment. To examine the influence of α on these partitioned heritability metrics, we first focus on the per-standardized annotation coefficients, as defined by Gazal et al. (2017) [6]:

$$\tau_c^* = \frac{M \cdot sd(a_c)}{h_g^2}\hat{\tau}_c \tag{5}$$



(a) Estimated $\tau^*(w)$

(b) Heritability enrichment

Figure 3: Comparison of meta-analyzed standardized coefficients and functional enrichment across different models of SNP heritability ($\alpha = 1$ on the x-axis and $\alpha = 0$ on the y-axis). The estimates are meta-analyzed over 47 GWAS traits using a random-effects model. Estimates are shown with standard errors. **(a)** shows standardized coefficients $\tau^*(w)$ and **(b)** shows estimated heritability enrichment.

The quantity τ_c^* is defined as additive change in per-SNP heritability associated to a 1 standard deviation increase in the value of the annotation $(sd(a_c))$, normalized by the average per-SNP heritability over all SNPs for the trait (h^2_{SNP}/M). Despite its usefulness for examining contributions to heritability across traits, the τ^* metric as defined above is not suitable for comparing models with different values of α, since the $\hat{\tau}_c$ in these models are on different scales (see Appendix B). Here, we propose a modified metric that captures the contributions of the different annotations to heritability as well as the influence of α:

$$\tau^*(w) = \frac{M \cdot sd(\text{Var}(X)^{1-\alpha} a_c)}{h_g^2} \hat{\tau}_c \tag{6}$$

The modified metric $\tau^*(w)$ has the same overall interpretation as the metric proposed by Gazal et al. (2017) [6], with the main difference being that we multiply the value of the annotation by the variance in allele frequency $\text{Var}(X)^{1-\alpha}$ when computing its standard deviation.

Figure 3(a) shows the estimates for the standardized coefficients for all the functional annotations that we analyzed, meta-analyzed across the 47 GWAS traits. We find that, in general, models with $\alpha = 0$ and $\alpha = 1$ produce comparable estimates of $\tau^*(w)$, with some notable exceptions (Supplementary Table S8). In particular, the models tend to diverge in the significance they assign to annotations associated with QTLs (e.g. the set of MaxCPP annotations introduced in [10], with the $\alpha = 0$ models often failing to reach the Bonferroni significance threshold for those annotations (3(a), Supplementary Table S8).

Comparable differences between the models are also observed for some of the LD-related annotations (e.g. CpG Content and Recombination Rate) that were introduced by Gazal et al. (2017) [6] (Figure 3(c), Supplementary Table S8). Note that some of these differences persist for intermediate values of α, though at a much smaller scale (Supplementary Figures 8-9). Thus the choice of α can influence the predicted biological relevance of different annotations even when frequency bins are used to model frequency-dependent effects.

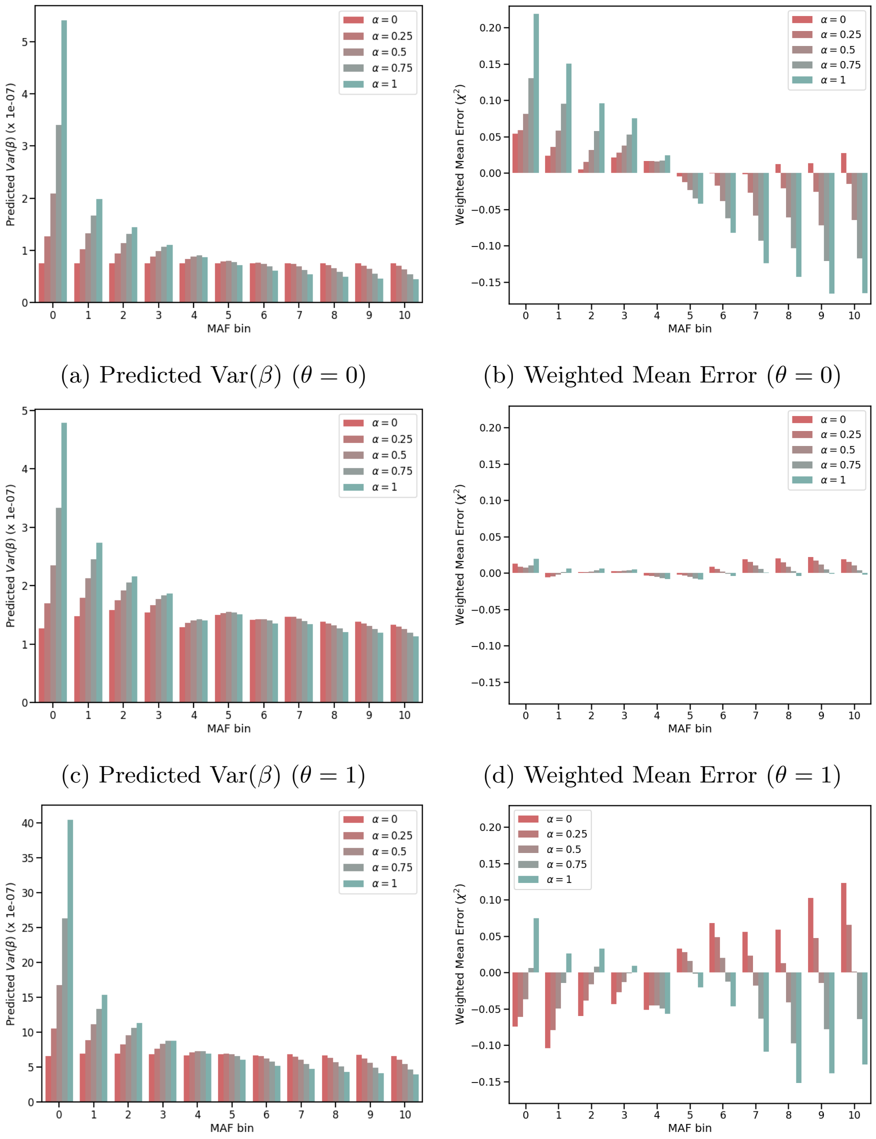
Figure 3(a) also has a slope higher than one, indicating that the choice $\alpha = 0$ leads to stronger estimates of enrichment for highly enriched categories. We see a similar effect when examining heritability enrichment, a measure of enrichment commonly-used for binary annotations. [5] (Figure 3(b), Supplementary Table S9). Heritability enrichment is the ratio of the proportion of heritability explained by SNPs in a given functional category to the proportion of SNPs in that category. As expected, intermediate values of α produce less systematic shift in the estimates of enrichment for both statistics (Supplementary Figure 10).

To test the significance of the enrichment estimates, we compute p-values using the differential enrichment metric as defined by Hujoel et al. (2019) [11] (see Appendix B). Even though $\alpha = 0$ produce higher enrichment results for strongly enriched categories, two functional categories that are deemed highly significant under the $\alpha = 1$ model fail to reach the significance threshold under the $\alpha = 0$ model (Enhancer (Hoffman) and TSS (Hoffman), Supplementary Table S9). For example, Enhancer (Hoffman) has unadjusted p-values of 0.229 under $\alpha = 0$ and 0.0005 under $\alpha = 1$. The opposite effect is observed for the Weak Enhancer (Hoffman) functional category, where the $\alpha = 0$ model reports it to be highly significant (Supplementary Table S9). These results again highlight the important role that the α parameter plays in the stratified LDSC framework, with different values of α potentially leading to different interpretations about the genetic architecture of complex traits.

3.4 The influence of α on models of SNP heritability

To explain the systematic shifts in enrichment as a function of α in the stratified models, here we empirically explore predicted mean squared effect size $\mathrm{Var}(\beta_j)$ and corresponding association statistic χ_j^2 as a function of allele frequency under the LDSC models.

In the standard univariate model ($\theta = 0$, $\alpha = 1$), rare variants are predicted to have very large effect relative to common variants (Figure 4(a)). As a result, the model tends to strongly overestimate the chi-square statistics for rare variants, and underestimate them for common variants (Figure 4(b)). This effect is more pronounced for $\alpha = 1$ than for any other choice of α (Figure 4(b)) and it helps explain the general trend of the standard univariate model producing downwardly biased estimates of SNP heritability [18]: to avoid having large squared errors for rare variants in the regression, the model must underestimate the effect sizes of common variants, thereby underestimating the total heritability. This bias can be reduced by choosing a more biologically plausible value of



(a) Predicted Var(β) ($\theta = 0$)

(b) Weighted Mean Error ($\theta = 0$)

(c) Predicted Var(β) ($\theta = 1$)

(d) Weighted Mean Error ($\theta = 1$)

(e) Predicted Var(β)
($\theta = 1$, HEC SNPs)

(f) Weighted Mean Error
($\theta = 1$, HEC SNPs)

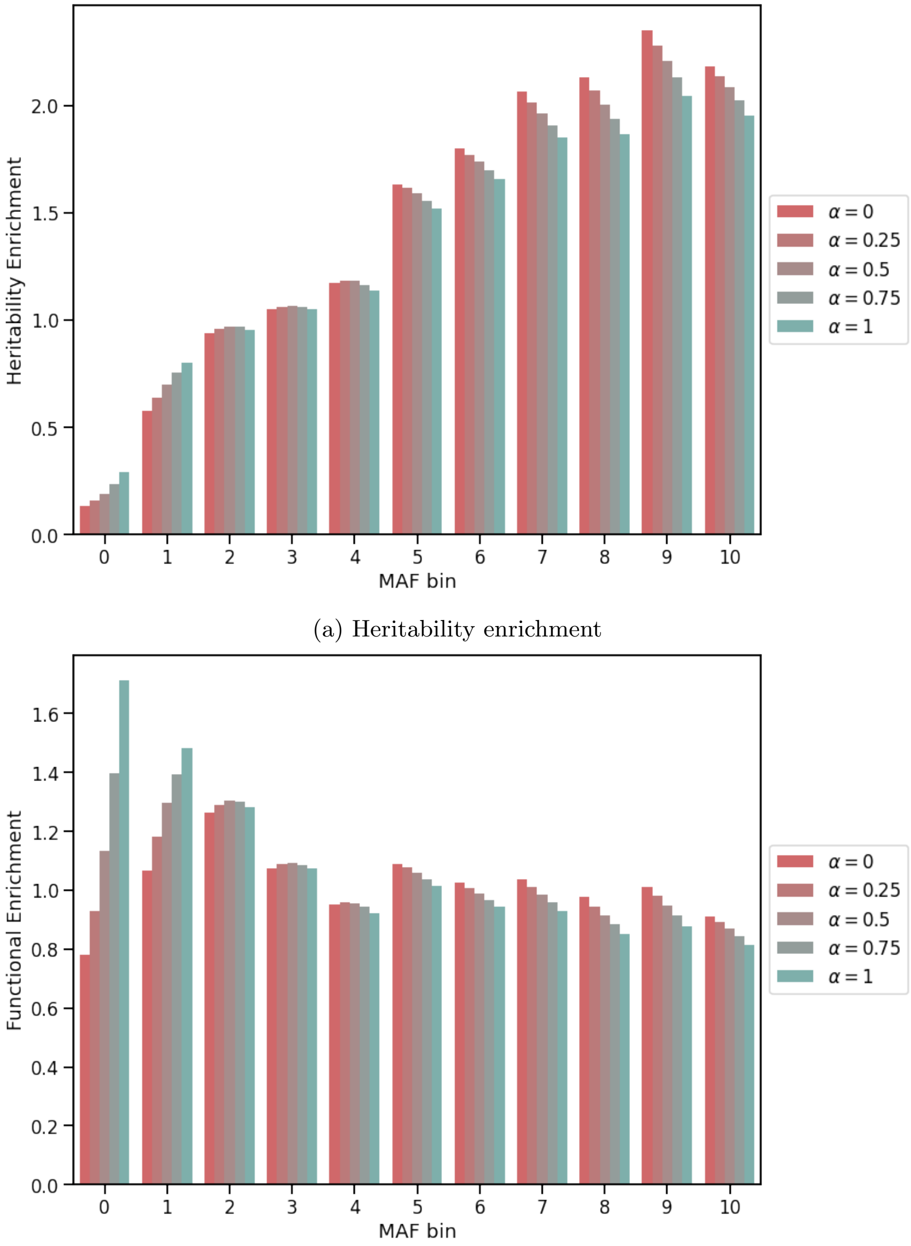
Figure 4: The predicted mean squared effect size and the mean error between predicted and observed χ^2 for SNPs in different MAF bins with different values of α. Estimates are averaged across 47 GWAS traits. **(a, c)** show predicted mean squared effect size for SNPs under **(a)** univariate models and **(c)** stratified models. **(b, d)** show mean error between predicted and observed χ^2 for SNPs under **(b)** univariate and **(d)** stratified models. Panels **(e, f)** show predicted mean squared effect size and mean error for SNPs in the top 10 Highly Enriched Categories (HECs) under the stratified models.

α, or by using stratified regression.

We expect the stratified models to better describe the relationship between allele frequency and effect size, since these models have the freedom to assign arbitrary mean predicted effect sizes for each frequency bin. Indeed, all stratified models correctly overcome the bias in predicting the χ^2 statistic (Figure 4(d)). However, they do so while predicting different effect sizes for rare and common variants (Figure 4(c)). Consequently, the weight of rare variants in the heritability of a trait can vary even in stratified models that agree on estimates of overall heritability. This, in turn, implies that the choice of α would lead to different proportions of heritability explained by each frequency bin and thus impact the predicted functional enrichment of variants in those bins (Figure 5(a)).

To explain the systematic differences in enrichment affecting highly enriched annotations, we now focus on the SNPs in the top 10 highly enriched functional categories (HECs, Supplementary Table S11), which comprise roughly 7% of all SNPs. The predicted mean effect size for rare functional variants is again very high for rare variants when $\alpha = 1$ (Figure 4(e)). And, as in the univariate models, this results in correspondingly high weighted mean errors in estimating the χ^2s for those SNPs (Figure 4(f)). This makes sense given the linear model for Var(β): for variants in strongly enriched categories, the annotation term overwhelms the frequency corrections and the distribution of effect sizes again follows the implicit assumptions induced by our choice of α (Figure 4(e)). Because the frequency bins fail to capture the frequency dependence for variants in highly enriched categories, the LDSC model for these variants is similar to a univariate model in that frequency effects are dictated by the choice of α. As in the univariate case, effects of rare variants are overestimated and, as argued above, the heritability contributed by these SNPs is underestimated. Figure 4 (b), (d) , and (f) also show that in all cases, intermediate α produce less biased predicted chi-squared, as measures by the weighted mean error, than either $\alpha = 0$ and $\alpha = 1$.

Finally, the choice of α also has a large impact on the inferred architecture of complex traits as measured by the proportion of heritability attributed to common and rare variants. Figure 5(a) shows markedly different estimates of the heritability enrichment metric among rare variants. Figure 5(a) also shows that heritability enrichment should not be used as a proxy for functional enrichment: Even though common variants have lower per-allele effect size (e.g., Figure 4c), they are highly enriched for heritability relative to expectations under the (unrealistic) $\alpha = 1$ model (Figure 5(a)). A better measure of functional enrichment would be measuring contributions to heritability relative to expectation under a constant per-allele effect size ($\alpha = 0$) model (see Appendix B). Figure 5(b) shows the estimated functional enrichment across frequency bins, revealing the expected behaviour of slight depletion for common variants and slight enrichment for rare variants, with substantial differences depending on α. The 0th bin shows the largest difference across choices of α, and also an unexpected functional depletion for the rarest variants and $\alpha = 0$. This 0th bin, which includes variants with allele frequency less than 5%, is often excluded



(a) Heritability enrichment

(b) Functional enrichment

Figure 5: Estimates of enrichment for the MAF bins for stratified LDSC model with five different values of α. Estimates are averaged across 47 GWAS traits. **(a)** shows the predicted heritability enrichment metric for 11 MAF bins and **(b)** shows the functional enrichment metric for the same categories.

from heritability estimates because of sensitivity to model assumptions and of other technical reasons [4].

4 Discussion

In this work, we proposed a reformulation of the LD Score Regression framework that replaces the r^2 measure of LD with the D^2 statistic, leveraging recently-proposed unbiased estimators [14]. The reformulation highlighted the implicit assumptions about the relationship between a variant's minor allele frequency and its effect size in commonly-used stratified LDSC models. This relationship, characterized by the parameter α, has been the subject of recent work on the frequency-dependent architecture of complex traits [18, 25, 15, 20].

In the initial LDSC model, this parameter was set to $\alpha = 1$ for mathematical convenience. Over the years, this assumption has been shown to be biologically implausible and to introduce biases in heritability estimates [18]. These biases have been primarily addressed in the literature through stratified LDSC, giving the model more parameters and thus more flexibility to overcome the unrealistic assumption. Despite the additional parameters, we have shown that the choice of α still has a substantial effect on partitioned heritability metrics, such as standardized heritability coefficients and functional enrichment.

These biases exist because the choice of $\alpha = 1$ implies rigid assumptions about the relationship between allele frequencies and effect sizes for variants in highly enriched categories. Under genome-wide empirical estimates of $\alpha = 0.38$ [15], the expected squared effect size of an uncommon variant at frequency $f = 0.05$ is 88% higher than a common variant at $f = 0.5$. Under the $\alpha = 1$ model, and for variants in highly enriched categories, the expected squared effect size of the uncommon variant is 526% that of common variant.

Even though such a synergistic effect is possible in theory, the analyses presented above and recent empirical estimates [20] suggest that values in the range $\alpha = 0.2 - 0.5$ are much more plausible across all functional annotations. Even though uncertainty remains about the 'best' choice for α, we find much more modest differences in inference results if α is chosen within this range. Therefore, to avoid letting the best be the enemy of the good, we recommend using the genome-wide average $\alpha = 0.38$ as a starting point for most stratified analyses.

Appendix A Derivation of LD Score Regression with the D^2 Statistic

Our re-formulation of the LD Score Regression model starts with a standard polygenic model where the phenotype of N individuals is assumed to linearly depend on M SNPs:

$$Y = X\beta + \epsilon \tag{7}$$

In this context, we assume that Y is a standardized vector of measured phenotypes and X is a $N \times M$ mean-centered (but not variance-normalized) genotype matrix. Following the derivation in Finucane et al. (2015) [5], we express the Ordinary Least Squares (OLS) solution for the marginal statistic of SNP j as a function of Y:

$$\begin{aligned}
\hat{\beta}_j &= \frac{X_j^\top Y}{X_j^\top X_j} = \frac{1}{X_j^\top X_j} X_j^\top \Big(X\beta + \epsilon \Big) \\
\hat{\beta}_j &= \frac{1}{X_j^\top X_j} \Big(\sum_k (X_j^\top X_k)\beta_k + X_j^\top \epsilon \Big) \\
\hat{\beta}_j &= \frac{1}{X_j^\top X_j} \Big(2N \sum_k D_{jk}\beta_k + X_j^\top \epsilon \Big)
\end{aligned} \tag{8}$$

Where the third line follows from the definition of the D statistic ($D_{jk} := \frac{1}{2N} X_j^\top X_k$). If we define the χ^2 association statistic for SNP j as $\chi_j^2 := N \left(X_j^\top X_j \right) \hat{\beta}_j^2$ and take the expectation over all the random components, we obtain:

$$\mathbb{E}[\chi_j^2] = \frac{4N^2}{X_j^\top X_j} \sum_k D_{jk}^2 \, \mathbb{E}[\beta_k^2] + \sigma_\epsilon^2 \tag{9}$$

Under this model, the total narrow-sense SNP heritability is defined as $h_{SNP}^2 = \sum_k \mathrm{Var}(X_k) \, \mathrm{Var}(\beta_k)$. Given this, coupled with the assumption that $h_g^2 + \sigma_\epsilon^2 = 1$, we obtain our general model for LD Score Regression with the D^2 statistic:

$$\mathbb{E}[\chi_j^2] = \frac{4N}{\mathrm{Var}(X_j)} \sum_k \mathrm{Var}(\beta_k) \Big(D_{jk}^2 - \frac{\mathrm{Var}(X_k)\,\mathrm{Var}(X_j)}{4N} \Big) + 1 \tag{10}$$

Assuming a large sample size for GWAS, this can be approximated by:

$$\mathbb{E}[\chi_j^2] \approx \frac{4N}{\mathrm{Var}(X_j)} \sum_k \mathrm{Var}(\beta_k) D_{jk}^2 + 1. \tag{11}$$

Finally, using the SNP heritability models outlined in Equation (2), the expression above can be equivalently formulated in terms of LD scores:

$$\begin{aligned}\mathbb{E}[\chi_j^2] &\approx N\Big[\tau_0 \ell(j,0) + \theta \sum_{c=1}^{96} \tau_c \ell(j,c)\Big] + 1 \\ \ell(j,c) &:= \frac{4}{\mathrm{Var}(X_j)} \sum_k D_{jk}^2 \,\mathrm{Var}(X_k)^{-\alpha} a_c(k).\end{aligned} \tag{12}$$

Appendix B Definitions of SNP Heritability and Functional Enrichment

B.1 SNP Heritability

SNP heritability quantifies the amount of phenotypic variance explained by a set of SNPs, genotyped or imputed [23]. If we assume that the phenotype is standardized and the genotype matrix is mean-centered (but not variance-normalized), we can write our model of SNP heritability as:

$$h_{SNP}^2 = \sum_k \mathrm{Var}(X_k)\,\mathrm{Var}(\beta_k) \tag{13}$$

Using the above formulation, the models of SNP heritability outlined in Equation (2) can be expressed as:

$$h_{SNP}^2 = \sum_k \mathrm{Var}(X_k)^{1-\alpha} \Big(\tau_0 + \theta \sum_{c=1}^{96} \tau_c a_c(k)\Big) \tag{14}$$

From the above expression, we can see that the relationship between partitioned heritability coefficients τ_c and the total SNP heritability h_{SNP}^2 depends on α. For instance, in the univariate case ($\theta = 0$), the coefficient τ_0 is proportional to per-SNP heritability and is related to the total SNP heritability as

$$\tau_0 = \frac{h_{SNP}^2}{\sum_k \mathrm{Var}(X_k)^{1-\alpha}} \tag{15}$$

When $\alpha = 1$, we recover the original LDSC formulation where the per-SNP heritability was assumed to be proportional to $1/M$ [4]. On the other hand, when $\alpha = 0$ the per-SNP heritability is inversely proportional to the average allele frequency variance across all SNPs.

B.2 Heritability and Functional Enrichment

Heritability enrichment is defined as the ratio of the proportion of heritability explained by SNPs in a given category c relative to the proportion of SNPs in that category [5]:

$$\text{Heritability Enrichment}(c) = \frac{h_{SNP}^2(c)/h_{SNP}^2}{M(c)/M} \tag{16}$$

Where $h^2_{SNP}(c)$ is the heritability explained by SNPs in category c and $M(c)$ is the number of SNPs in that category. To test for significance of the functional enrichment metric, we use the differential enrichment metric as defined in Hujoel et al. (2019) [11]:

$$\text{Differential Enrichment}(c) = \frac{h^2_{SNP}(c)}{M(c)} - \frac{h^2_{SNP} - h^2_{SNP}(c)}{M - M(c)} \tag{17}$$

As outlined in the main text (Section 2.5), a better measure of functional enrichment would be measuring contributions to heritability relative to expectation under a constant per-allele effect size ($\alpha = 0$) model. This implies that instead of dividing by the proportion of SNPs, we divide by the proportion of genotypic variances:

$$\text{Functional Enrichment}(c) = \frac{h^2_{SNP}(c)/h^2_{SNP}}{\sum_{j\in c} Var(X_j)/\sum_j Var(X_j)} \tag{18}$$

Appendix C Analysis Procedures and Evaluation Metrics

C.1 Computing and Comparing LD Scores from 1000 Genomes Data

To conduct the analyses discussed in this paper, we computed LD scores using three different estimators of LD (naive r^2, corrected r^2 [4, 24] and D^2 [14]) for $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and for 3 of the super populations in Phase III of the 1000 Genomes Project (Africans, Asians, and Europeans) [1, 21], excluding the sets of closely related individuals identified in [9]. We followed the same quality control procedures as in the original LDSC software documentation [4] (e.g. discarding singleton SNPs). All of these quality control steps were run with `PLINK v1.9` [13].

To qualitatively compare LD Scores from the univariate model with different values of α, we normalized them such that the slope of the univariate regression becomes $\frac{N}{M}h^2_{SNP}$ instead of $N\tau_0$. Given our definition of τ_0 in terms of h^2_{SNP} in Equation (15), we see that this can be achieved by dividing the LD score for each model by the following quantity:

$$\frac{1}{M}\sum_k \text{Var}(X_k)^{1-\alpha}$$

By construction, when $\alpha = 1$, as in the standard LDSC models, the normalization factor is 1. When $\alpha = 0$, the normalization factor becomes the average genome-wide variance in allele frequency.

C.2 Parameter Estimation and Meta-analysis

To estimate the parameters of the various heritability models, we used the Iteratively re-weighted least squares (IRLS) jackknife estimator from Bulik-Sullivan et al. (2015) [4] with some modifications to account for differences in the heritability models. Primarily, since our general model starts with an unnormalized genotype matrix, the total SNP heritability is given by:

$$h^2_{SNP} = \tau_0 \sum_k \mathrm{Var}(X_k)^{1-\alpha} + \theta \sum_{c=1}^{96} \tau_c \sum_k \mathrm{Var}(X_k)^{1-\alpha} a_c(k)$$

Once the coefficients $\hat{\tau}_c$ have been estimated by the `ldsc` software, to obtain the total heritability, we have to multiply them by different factors. For example, in the case of the $\alpha = 1$ models, we multiply by the sum of the annotations $\sum_j a_c(k)$, whereas for the $\alpha = 0$ models, we multiply by the sum of the annotations weighted by MAF: $\sum_k \mathrm{Var}(X_k) a_c(k)$.

To meta-analyze the coefficients as well as measures of functional enrichment across the 47 GWAS traits, we used a random effects model as implemented in the `R` package `meta`[2].

C.3 Empirical Evaluation of SNP heritability models

To empirically evaluate the stratified SNP heritability models as in Section 2.5, we fit each model to the summary statistics from 47 GWAS traits (Supplementary Table S1) using the `ldsc` software [4] with standard configurations. For each trait and model, we obtained the mean parameter estimates $\hat{\tau}_c$ and used them to compute the predicted $\mathbb{E}[\hat{\chi}^2_j]$ for all SNPs that were used in the regression. Then, for each MAF bin, we computed the weighted mean error using the following equation:

$$\text{Weighted Mean Error} = \frac{1}{\sum_{j \in \text{MAFbin}} w_j} \sum_{j \in \text{MAFbin}} w_j (\mathbb{E}[\hat{\chi}^2_j] - \chi^2_j) \qquad (19)$$

Where the $\mathbb{E}[\hat{\chi}^2_j]$ is the predicted association statistic under the model and χ^2_j is the observed statistic. The weight w_j is simply the LD score weight that we employed in the regression [4]. This strategy of weighing the performance statistics by the LD score weights has been explored in previous work [20]. Since, in our case, each model has its own set of LD score weights that match its α value, here we used the LD score weights for the $\alpha = 1$ model throughout. The weighted mean errors per MAF bin that we report in the main text are averaged across the 47 GWAS traits.

Supplemental Data

Supplemental Data include ten figures (Supplementary Figures S1-10) and eleven tables (Supplementary Tables S1-11).

Declaration of Interests

The authors declare no conflict of interest.

Acknowledgements

We thank Chris Gignoux and members of the Gravel Lab for useful discussions.

Web Resources

Baseline-LD model version 2.2,
`data.broadinstitute.org/alkesgroup/LDSCORE/`
Summary statistics for 47 GWAS traits,
`data.broadinstitute.org/alkesgroup/LDSCORE/independent_sumstats/`

Data and Code Availability

Code to compute the LD scores with the unbiased estimator of D^2 and carry out the analyses discussed in this paper is available on github:
`https://github.com/shz9/unbiased-ldsc`.

References

[1] David L. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, Francisco M. De La Vega, Peter Donnelly, Michael Egholm, et al. "A map of human genome variation from population-scale sequencing". In: *Nature* (2010). ISSN: 14764687. DOI: 10.1038/nature09534.

[2] Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. "How to perform a meta-analysis with R: a practical tutorial". In: *Evidence-Based Mental Health* 22 (2019), pp. 153–160.

[3] Brendan Bulik-Sullivan, Hilary K. Finucane, Verneri Anttila, Alexander Gusev, Felix R. Day, Po Ru Loh, Laramie Duncan, John R.B. Perry, Nick Patterson, Elise B. Robinson, et al. "An atlas of genetic correlations across human diseases and traits". In: *Nature Genetics* (2015). ISSN: 15461718. DOI: 10.1038/ng.3406.

[4] Brendan Bulik-Sullivan, Po Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, Benjamin M. Neale, Aiden Corvin, et al. "LD score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* (2015). ISSN: 15461718. DOI: 10.1038/ng.3211.

[5] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nature Genetics* (2015). ISSN: 15461718. DOI: 10.1038/ng.3404.

[6] Steven Gazal, Hilary K. Finucane, Nicholas A. Furlotte, Po Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M. Neale, Alexander Gusev, et al. "Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection". In: *Nature Genetics* (2017). ISSN: 15461718. DOI: 10.1038/ng.3954.

[7] Steven Gazal, Po Ru Loh, Hilary K. Finucane, Andrea Ganna, Armin Schoech, Shamil Sunyaev, and Alkes L. Price. "Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations". In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-018-0231-8.

[8] Steven Gazal, Carla Marquez-Luna, Hilary K. Finucane, and Alkes L. Price. *Reconciling S-LDSC and LDAK functional enrichment estimates.* 2019. DOI: 10.1038/s41588-019-0464-1.

[9] Steven Gazal, Mourad Sahbatou, Marie Claude Babron, Emmanuelle Genin, and Anne Louise Leutenegger. "High level of inbreeding in final phase of 1000 Genomes Project". In: *Scientific Reports* (2015). ISSN: 20452322. DOI: 10.1038/srep17453.

[10] Farhad Hormozdiari, Steven Gazal, Bryce Van De Geijn, Hilary K. Finucane, Chelsea J.T. Ju, Po Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'connor, et al. "Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits". In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-018-0148-2.

[11] Margaux L.A. Hujoel, Steven Gazal, Farhad Hormozdiari, Bryce van de Geijn, and Alkes L. Price. "Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species". In: *American Journal of Human Genetics* (2019). ISSN: 15376605. DOI: 10.1016/j.ajhg.2019.02.008.

[12] Bogdan Pasaniuc and Alkes L. Price. *Dissecting the genetics of complex traits using summary association statistics.* 2017. DOI: 10.1038/nrg.2016.142.

[13] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. De Bakker, Mark J. Daly, et al. "PLINK: A tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* (2007). ISSN: 00029297. DOI: 10.1086/519795.

[14] Aaron P. Ragsdale and Simon Gravel. "Unbiased Estimation of Linkage Disequilibrium from Unphased Data". In: *Molecular Biology and Evolution* (2020). ISSN: 15371719. DOI: 10.1093/molbev/msz265.

[15] Armin P. Schoech, Daniel M. Jordan, Po Ru Loh, Steven Gazal, Luke J. O'Connor, Daniel J. Balick, Pier F. Palamara, Hilary K. Finucane, Shamil R. Sunyaev, and Alkes L. Price. "Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection". In: *Nature Communications* (2019). ISSN: 20411723. DOI: 10.1038/s41467-019-08424-6.

[16] Huwenbo Shi, Steven Gazal, Masahiro Kanai, Evan M Koch, Armin P Schoech, Samuel S Kim, Yang Luo, Tiffany Amariuta, Yukinori Okada, Soumya Raychaudhuri, et al. "Population-specific causal disease effect sizes in functionally important regions impacted by selection". In: *bioRxiv* (2019). DOI: 10.1101/803452.

[17] Doug Speed and David J. Balding. "SumHer better estimates the SNP heritability of complex traits from summary statistics". In: *Nature Genetics* (2019). ISSN: 15461718. DOI: 10.1038/s41588-018-0279-5.

[18] Doug Speed, Na Cai, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. "Reevaluation of SNP heritability in complex human traits". In: *Nature Genetics* (2017). ISSN: 15461718. DOI: 10.1038/ng.3865.

[19] Doug Speed, Gibran Hemani, Michael R. Johnson, and David J. Balding. "Improved heritability estimation from genome-wide SNPs". In: *American Journal of Human Genetics* (2012). ISSN: 00029297. DOI: 10.1016/j.ajhg.2012.10.010.

[20] Doug Speed, John Holmes, and David J. Balding. "Evaluating and improving heritability models using summary statistics". In: *Nature Genetics* (2020). ISSN: 15461718. DOI: 10.1038/s41588-020-0600-y.

[21] The 1000 Genomes Project Consortium. "An integrated map of genetic variation". In: *Nature* (2012).

[22] Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, et al. "Common SNPs explain a large proportion of the heritability for human height". In: *Nature Genetics* (2010). ISSN: 10614036. DOI: 10.1038/ng.608.

[23] Jian Yang, Jian Zeng, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. "Concepts, estimation and interpretation of SNP-based heritability". In: (2017). ISSN: 15461718. DOI: 10.1038/ng.3941.

[24] Ping Yin and Xitao Fan. "Estimating r2 shrinkage in multiple regression: A comparison of different analytical methods". In: *Journal of Experimental Education* (2001). ISSN: 19400683. DOI: 10.1080/00220970109600656.

[25] Jian Zeng, Ronald De Vlaming, Yang Wu, Matthew R. Robinson, Luke R. Lloyd-Jones, Loic Yengo, Chloe X. Yap, Angli Xue, Julia Sidorenko, Allan F. McRae, et al. "Signatures of negative selection in the genetic architecture of human complex traits". In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-018-0101-4.