Causal considerations can determine the utility of machine learning assisted GWAS

Sumit Mukherjee*[1], Zachary McCaw*[1], David Amar[1], Rounak Dey[1], Thomas Soare[1], Kaiwen Xu[1], Hari Somineni[1], insitro Research Team, Nicholas Eriksson[1], Colm O'Dushlaine[1]

Abstract

Machine Learning (ML) is increasingly employed to generate phenotypes for genetic discovery, either by imputing existing phenotypes into larger cohorts or by creating novel phenotypes. While these ML-derived phenotypes can significantly increase sample size, and thereby empower genetic discovery, they can also inflate the false discovery rate (FDR). Recent research has focused on developing estimators that leverage both true and machine-learned phenotypes to properly control the type-I error. Our work complements these efforts by exploring how the true positive rate (TPR) and FDR depend on the causal relationships among the inputs to the ML model, the true phenotypes, and the environment.

Using a simulation-based framework, we study architectures in which the machine-learned proxy phenotype is derived from biomarkers (i.e. inputs) either causally upstream or downstream of the target phenotype. We show that no inflation of the false discovery rate occurs when the proxy phenotype is generated from upstream biomarkers, but that false discoveries can occur when the proxy phenotype is generated from downstream biomarkers. Next, we show that power to detect variants truly associated with the target phenotype depends on its heritability and correlation with the proxy phenotype. However, the source of the correlation is key to evaluating a proxy phenotype's utility for genetic discovery. We demonstrate that evaluating machine-learned proxy phenotypes using out-of-sample predictive performance (e.g. phenotypic correlation) provides a poor lens on utility. This is because overall predictive performance does not differentiate between genetic and environmental correlation. In addition to parsing these properties of machine-learned phenotypes via simulations, we further illustrate them using real-world data from the UK Biobank.

Keywords: Machine Learning derived phenotype, GWAS, Imputation, Prediction-based Inference

[1]insitro inc., South San Francisco, California, USA.

*Equal contributing authors listed in random order.

1 Introduction

Genome-wide association studies (GWAS) identify genetic variations that are associated with a particular phenotype, and have revolutionized our understanding of the genetic architecture of complex traits and diseases [1]. This approach has successfully uncovered numerous genetic variants that contribute to the risk of complex disorders, paving the way for precision medicine and targeted therapeutic interventions. The advent of large-scale biobanks has further propelled the field of genetic discovery by providing structured, well-powered, data sets with genotypes and deep phenotyping for hundreds of thousands of individuals [2, 3, 4, 5]. The quality and unprecedented scale of these data sets have empowered researchers to detect genetic variants with ever smaller effect sizes while capturing an ever greater proportion of complex trait heritability [6].

Despite the size and scope of population-based biobanks, the challenge of data sparsity persists. Many phenotypes of interest are measured for only a subset of participants, and only a fraction of participants will develop any given disease, limiting the effective sample size available for GWAS in observational biobanks. This sparsity can meaningfully diminish the power to detect associations for complex traits with modest genetic effects. To address this limitation, researchers increasingly turn to machine learning (ML) to impute missing phenotypic values from the available data. These ML imputation approaches have been shown to improve genetic discovery for difficult to ascertain phenotypes, such as the optical cup-to-disc ratio, thoracic aortic diameter, major depression, and hepatic fat percentage [7, 8, 9, 10, 11, 12, 10].

When performing GWAS on predicted or imputed outcomes, the relationship between genotype and the imputed phenotype may differ both quantitatively and qualitatively (i.e. in existence or direction) from that between genotype and the target outcome [13, 14, 15]. Distortion of the genotype-phenotype relationship due to imputation can lead to inflated type I error, due to the detection of signals that spuriously associate with the imputed phenotype but not the true phenotype, and compromise the utility of downstream analyses that depend on unbiased effect size estimation, such as polygenic scoring. Recent work [16, 17, 13, 14, 15] has focused on developing methods to address these challenges and minimize the risk of false discoveries. Specifically, these methods aim to provide unbiased estimation for the effect of genotype on the target outcome that is robust to the accuracy or quality of the imputation model. In doing so, these methods guarantee valid inference, but may forego power as compared with the simpler strategy of proxy GWAS (**Figure 1**), where the machine-learned proxy phenotype is studied in place of the original phenotype.

This work is intended to complement the recent work in prediction-based inference by studying how the power and false discovery rate of proxy GWAS depend on the causal relationships among genotypes, the true phenotype, and the variables the enter the imputation model, which we describe as biomarkers. The remainder of this paper is organized as follows. Section 2 describes our simulation framework and details of the real data analysis. Section 3 includes our case-studies on simulated and real data. We first contrast proxy phenotypes imputed from upstream versus downstream biomarkers in terms of their power for detecting

true positive association and their FDR. Second, we examine how heritability of the true phenotype and its correlation with the proxy phenotype affect power to recover true positive associations. Third, we decompose the phenotypic correlation, assessing the dependence of the true positive rate (TPR) and the FDR on genetic versus environmental correlations. We conclude with the implications of our case-studies for practice.

2 Data and Methods

2.1 Real-world dataset

The UK Biobank (UKB) is a large-scale prospective cohort study and biomedical database containing detailed genetic and health information on approximately 500,000 individuals from the United Kingdom, aged between 40 and 69 years at the time of recruitment [18]. The UKB resource encompasses a wide variety of data on health-related outcomes, including hospital records, cancer registries, death records, and physical measurements, as well as self-reported health questionnaires. Additionally, the resource contains a vast array of biological measurements such as blood, urine, and saliva biochemistry.

2.2 Methods for simulated traits

2.2.1 Simulating traits with a specified heritability and genetic correlation

Let $\mathbf{G}$ denote a vector of J genetic variants in linkage equilibrium, with elements $G_j \sim \text{Binom}(2, p)$, where p is the minor allele frequency. $\boldsymbol{G}$ can contain both causal and non-causal variants for biomarkers and phenotypes. Let S_{bio} and S_{pheno} denote the indices of the causal variants for biomarkers and phenotypes. Non-causal variants have effect sizes of zero and are therefore not included in either S_{bio} or S_{pheno}. The sets S_{bio} and S_{pheno} may or may not overlap.

For the jth variant causal for biomarkers, let $\boldsymbol{\beta}_{\text{bio},j}$ denote an $n_{\text{bio}} \times 1$ vector representing the non-zero effects of the jth variant on the n_{bio} biomarkers. Similarly, let $\boldsymbol{\beta}_{\text{pheno},j}$ denote the $n_{\text{pheno}} \times 1$ vector of non-zero effects for the jth variant in S_{pheno} on the n_{pheno} phenotypes. We assume that for each causal variant, the vectors $\boldsymbol{\beta}_{\text{bio},j}$ and $\boldsymbol{\beta}_{\text{pheno},j}$ follow a joint multivariate normal distribution:

$$\begin{pmatrix} \boldsymbol{\beta}_{\text{bio},j} \\ \boldsymbol{\beta}_{\text{pheno},j} \end{pmatrix} \sim \mathcal{N}\left\{ \mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}_{\text{bio}} & \boldsymbol{\Sigma}_{\text{bio,pheno}} \\ \boldsymbol{\Sigma}_{\text{bio,pheno}}^{\top} & \boldsymbol{\Sigma}_{\text{pheno}} \end{pmatrix} \right\}. \tag{1}$$

Here, $\boldsymbol{\Sigma}_{\text{bio}}$ is an $n_{\text{bio}} \times n_{\text{bio}}$ diagonal matrix with the variances of the effect sizes for each biomarker on the diagonal. The diagonal elements $\Sigma_{\text{bio},kk} = \sigma^2_{\text{bio},k}$ determine the contribution of the jth genetic variant to the heritability of the kth biomarker. Likewise, $\boldsymbol{\Sigma}_{\text{pheno}}$ is a $n_{\text{pheno}} \times n_{\text{pheno}}$ diagonal matrix with the effect size variances for each phenotype on the diagonal. The diagonal elements $\Sigma_{\text{pheno},kk} = \sigma^2_{\text{pheno},k}$ determine the contribution of the genetic variance to the heritability of the kth phenotype. Lastly, $\boldsymbol{\Sigma}_{\text{bio,pheno}}$ is the $n_{\text{bio}} \times n_{\text{pheno}}$ matrix where each element $\Sigma_{\text{bio,pheno},kk'}$ represents the covariance between the effect sizes of

the kth biomarker and the k'th phenotype due to the jth causal variant.

The final $n_{\text{bio}} \times 1$ biomarker vector and $n_{\text{pheno}} \times 1$ phenotype vector are generated as:

$$\mathbf{b} = \sum_{j \in S_{\text{bio}}} G_j \boldsymbol{\beta}_{\text{bio},j} + \boldsymbol{\epsilon}_{\text{bio}}, \tag{2}$$

$$\mathbf{p} = \sum_{j \in S_{\text{pheno}}} G_j \boldsymbol{\beta}_{\text{pheno},j} + \boldsymbol{\epsilon}_{\text{pheno}}. \tag{3}$$

The environmental components are simulated as:

$$\begin{pmatrix} \boldsymbol{\epsilon}_{\text{bio}} \\ \boldsymbol{\epsilon}_{\text{pheno}} \end{pmatrix} \sim \mathcal{N}\left\{ \mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}_{\epsilon,\text{bio}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\epsilon,\text{pheno}} \end{pmatrix} \right\}.$$

Here $\boldsymbol{\Sigma}_{\epsilon,\text{bio}}$ and $\boldsymbol{\Sigma}_{\epsilon,\text{pheno}}$ are diagonal $n_{\text{bio}} \times n_{\text{bio}}$ and $n_{\text{pheno}} \times n_{\text{pheno}}$ matrices defining the environmental contributions to the variances of biomarkers and phenotypes respectively.

2.2.2 Generating downstream traits with specified degree of environmental influence and direct genetic effects

When generating downstream phenotypes $\boldsymbol{p}_{\text{down}}$, we first simulate upstream biomarkers $\boldsymbol{b}_{\text{up}}$ following (2), then construct the phenotypes as follows:

$$\mathbf{p}_{\text{down}} = \mathbf{A}_{\text{bio}} \mathbf{b}_{\text{up}} + \sum_{j \in S_{\text{pheno,down}}} G_j \boldsymbol{\beta}_{\text{pheno,down},j} + \boldsymbol{\epsilon}_{\text{pheno,down}} \tag{4}$$

Here $\boldsymbol{A}_{\text{bio}}$ is an $n_{\text{pheno}} \times n_{\text{bio}}$ matrix representing a linear relationship between upstream biomarkers and downstream phenotypes, $S_{\text{pheno,down}}$ is the set of indices for the variants in $\boldsymbol{G}$ with direct causal effects on the downstream phenotype, $\boldsymbol{\beta}_{\text{pheno,down},j}$ is an $n_{\text{pheno}} \times 1$ vector of random effect sizes drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{pheno,down}})$ distribution, where $\boldsymbol{\Sigma}_{\text{pheno,down}}$ is a diagonal matrix, and $\boldsymbol{\epsilon}_{\text{pheno,down}}$ is an $n_{\text{pheno}} \times 1$ residual drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon,\text{down}})$ distribution, where $\boldsymbol{\Sigma}_{\epsilon,\text{down}}$ is again a diagonal matrix. The diagonal elements of $\boldsymbol{\Sigma}_{\text{pheno,down}}$ determine the contribution of direct genetic effects to the variance of $\boldsymbol{p}_{\text{down}}$, while those of $\boldsymbol{\Sigma}_{\epsilon,\text{down}}$ determine the contribution of environmental effects.

Analogously, to generate downstream biomarkers $\boldsymbol{b}_{\text{down}}$, we first simulate upstream phenotypes $\boldsymbol{p}_{\text{up}}$ following (3), then construct the biomarkers as follows:

$$\mathbf{b}_{\text{down}} = \mathbf{A}_{\text{pheno}} \mathbf{p}_{\text{up}} + \sum_{j \in S_{\text{bio,down}}} G_j \boldsymbol{\beta}_{\text{bio,down},j} + \boldsymbol{\epsilon}_{\text{bio,down}} \tag{5}$$

Here $\boldsymbol{A}_{\text{pheno}}$ is an $n_{\text{bio}} \times n_{\text{pheno}}$ matrix representing a linear relationship between upstream phenotypes and downstream biomarkers, $S_{\text{bio,down}}$ is the set of indices for the variants in $\boldsymbol{G}$ with direct causal effects on the downstream biomarkers, $\boldsymbol{\beta}_{\text{bio,down},j}$ is an $n_{\text{bio}} \times 1$ vector of random effect sizes drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{bio,down}})$ distribution, and $\boldsymbol{\epsilon}_{\text{bio,down}}$ is an $n_{\text{bio}} \times 1$ residual drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon,\text{down}})$ distribution.

2.2.3 Heritability and Genetic Correlation Estimation

The heritability for each biomarker and phenotype was estimated as the proportion of the phenotypic variance explained by the genetic variants. For a set of N individuals with J genetic variants, the genotype matrix is denoted by $\mathbf{G} \in \mathbb{R}^{N \times J}$. Let $\boldsymbol{\beta}_{\text{bio}} \in \mathbb{R}^{n_{\text{bio}} \times J}$ and $\boldsymbol{\beta}_{\text{pheno}} \in \mathbb{R}^{n_{\text{pheno}} \times J}$ be the matrices of genetic effects for n_{bio} biomarkers and n_{pheno} phenotypes respectively. The heritability of biomarkers (h^2_{bio}) and phenotypes (h^2_{pheno}) are calculated as follows:

$$h^2_{\text{bio}} = \frac{\text{Var}\left(\mathbf{G}\boldsymbol{\beta}^\top_{\text{bio}}\right)}{\text{Var}\left(\mathbf{B}\right)}, \quad h^2_{\text{pheno}} = \frac{\text{Var}\left(\mathbf{G}\boldsymbol{\beta}^\top_{\text{pheno}}\right)}{\text{Var}\left(\mathbf{P}\right)}, \tag{6}$$

where $\mathbf{B} \in \mathbb{R}^{N \times n_{\text{bio}}}$ and $\mathbf{P} \in \mathbb{R}^{N \times n_{\text{pheno}}}$ are the matrices of biomarker values and phenotypic values, respectively.

The genetic correlation among biomarkers ($\boldsymbol{\rho}_{\mathbf{B}}$), among phenotypes ($\boldsymbol{\rho}_{\mathbf{P}}$), and between biomarkers and phenotypes ($\boldsymbol{\rho}_{\mathbf{BP}}$) for shared genetic variants were calculated as:

$$\boldsymbol{\rho}_{\text{bio}} = \text{Corr}\big(\boldsymbol{\beta}_{\text{bio}}, \boldsymbol{\beta}_{\text{bio}}\big), \quad \boldsymbol{\rho}_{\text{pheno}} = \text{Corr}\big(\boldsymbol{\beta}_{\text{pheno}}, \boldsymbol{\beta}_{\text{pheno}}\big), \quad \boldsymbol{\rho}_{\text{bio,pheno}} = \text{Corr}(\boldsymbol{\beta}_{\text{bio}}, \boldsymbol{\beta}_{\text{pheno}}),$$

where, for example, the correlation is calculated as:

$$\boldsymbol{\rho}_{\text{bio,pheno},kk'} = \text{Corr}(\boldsymbol{\beta}_{\text{bio},k}, \boldsymbol{\beta}_{\text{pheno},k'})$$

2.2.4 Association testing for simulated traits

Genome-wide association testing for simulated traits was conducted using per-variant linear regression analyses. Specifically, the kth biomarker $\boldsymbol{b}_k$ or phenotype $\boldsymbol{p}_k$ was associated with the jth column of the genotype matrix $\boldsymbol{G}_j$ according to the model:

$$\mathbb{E}(\boldsymbol{y}_k | \boldsymbol{G}_j) = \alpha_0 + \alpha_G \boldsymbol{G}_j,$$

where $\boldsymbol{y}_k \in \mathbb{R}^{N \times 1}$ is either $\boldsymbol{b}_k$ or $\boldsymbol{p}_k$, α_0 is an intercept, α_G is the genetic effect. For each $\boldsymbol{G}_j$, representing a single genetic variant, the null hypothesis $H_0 : \alpha_G = 0$ was evaluated via a standard Wald test. Genome-wide significance was declared at the Bonferroni threshold of $0.05/J$, where J is the total number of variants tested for association.

Across a simulated GWAS, the true positive rate (TPR) was defined as the ratio of the number of causal variants that reached genome-wide significance to the total number of causal variants, The false discovery (FDR) was defined as ratio of the number of non-causal variants that reached genome-wide significance to the total number of variants that reach genome-wide significance.

2.3 Methods for real data experiments

2.3.1 Genetic data processing and analyses

To avoid confounding due to population structure, the UK Biobank (UKB) was subset to unrelated subjects of White-British ancestry [19, 18]. Imputed genotypes were filtered to those having a minor allele frequency $> 1\%$, INFO score > 0.8, and Hardy-Weinberg equilibrium $P > 1 \times 10^{-10}$. The following standard covariates were included in all GWAS: age, sex, genotyping array, and the top 20 genetic principal components [18]. Genome-wide association studies (GWAS) and clumping were performed using `PLINK` (v1.9) [20]. GWAS for quantitative traits were performed with linear regression models, and GWAS for binary traits with logistic regression models. Genetic correlation between two traits was estimated using `LDSC v1.0.1` with default settings [21].

2.3.2 Phenotype preparation

Height (UKB: 50), weight (UKB: 21002), and circulating urate (UKB: 30880) were obtained directly from the UKB, filtered non-missing values, and rank-normal transformed [22]. As algorithmically-defined gout was not directly available from UKB, a gout phenotype was constructed following [23]. Specifically, a patient was labeled as having gout if they satisfied at least one of: (1) self-reported gout (code: 1466; UKB: 20002), (2) had an ICD10 code for gout (code: M10; UKB: 41202, 41204, 41270), (3) reported taking allopurinol (1140875408), sulfinpyrazone (1140909890), or colchicine (1140875486) in field 20003, and (4) did not have a hospital diagnosis of leukaemia or lymphoma (codes: C81–C96).

2.3.3 Generating proxy phenotypes with specified target-phenotype correlation

A noisy phenotype Y_ρ having specified correlation ρ with a target phenotype Y can be generated via:

$$Y_\rho = \rho \cdot Y + \sqrt{1-\rho^2} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{7}$$

Here $\rho \in [0,1]$ controls the expected correlation between Y_ϵ and Y, and ϵ is mean-zero noise generated independently of Y.

2.3.4 Generating a purely environmental proxy phenotype

To construct an imputed phenotype with minimal autosomal heritability, a linear model was fit to predict circulating creatinine (UKB: 30700), a byproduct of protein catabolism, on the basis of age (UKB: 21022), genetic sex (UKB: 22001), and daily alcohol intake, computed as in [24]. The model inputs included polynomial features up to 2nd degree, including all quadratic terms and pairwise interactions:

$$\text{Creatinine} \sim (\text{Age}, \text{Sex}, \text{Alcohol Intake})^{\otimes 2}$$

To modulate their association with measured creatinine Y, the predicted creatinine levels $\widehat{Y}$ were standardized to have mean 0 and unit variance, then corrupted by introducing noise as

in (7).

3 Results

3.1 Downstream vs. Upstream Biomarkers for imputation - which to use?

We first explored how the causal relationship of biomarkers to the true phenotype affects genetic discovery when a machine-learned proxy phenotype, imputed from biomarkers, is studied in place of the true phenotype. **Figure 2** depicts the causal diagrams for the different data generating processes studied, along with the empirical true positive rate (TPR) and the false discovery rate (FDR) as a function of the heritability of the true phenotype P and the mean heritability of the biomarkers. We consider four scenarios, in each case using GWAS on the proxy phenotype P^* as a means of identifying genetic variants associated with the true phenotype P. In scenarios (A) and (B), the biomarkers B utilized to generate the proxy phenotype lie upstream of the true phenotype on the causal pathway. In (A), the effect of genotype on the true phenotype is fully mediated by the biomarkers, whereas in (B) genotype affects P both directly and indirectly via B. In both (A) and (B), the FDR is zero. This is because all variants causal for the biomarkers are in fact causal for the true phenotype when B lies upstream P on the causal pathway. However, scenarios (A) and (B) differ with respect to the TPR. When the effect of genotype is fully mediated by biomarkers as in (A), all variants causal for true phenotype are also causal for the biomarkers. Thus, by studying a proxy phenotype P^* that is a composite of B, it should be possible to recover all variants causal for P. As the sample size and biomarker heritability increase, the TPR in scenario (A) will approach 1. In contrast, when the effect of genotype is only partially mediated by the biomarkers as in (B), there exist variants with effects on P that do not have effects on B. Even with increasing sample size and biomarker heritability, it is not expected that variants whose effects on P are not mediated by B can be detected by studying P^*. In general, the TPR in (B) will be bounded above by the fraction of variants causal for the true phenotype whose effects are mediated by the biomarkers included in the proxy phenotype.

In scenarios (C) and (D), the biomarkers B utilized to generate the proxy phenotype lie downstream of the true phenotype on the causal pathway. In (C), the effect of genotype on the biomarkers is fully mediated by the true phenotype, while in (D) G affects B both directly and via P. In both (C) and (D) the TPR is high, and will approach 1 as the sample size and biomarker heritability increase. This is because all variants causal for the true phenotype are ultimately causal for the collection of biomarkers. Thus, by studying a proxy phenotype derived from the downstream biomarkers, all variants causal for P can be detected. Where (C) and (D) differ is with respect to the FDR. In (C), there are no variants with effects on B that do not have effects on P. Consequently, the FDR is again zero. However, when the effect of genotype on the biomarkers is only partially mediated by the true phenotype as in (D), there exist variants with effects on B that do not have effects on P. Such variants, which are not relevant to P, are expected to surface in GWAS of P^* as sample size and biomarker heritability increase. In general, the FDR in (D) will be bounded

below by the fraction of variants causal for the biomarkers included in the proxy phenotype whose effects are not mediated by the true phenotype.

To illustrate the ideas examined by these simulations, we considered GWAS of two pairs of traits whose causal relationship is well established: Urate $\rightarrow$ Gout and Height $\rightarrow$ Weight. For clarity, we consider the simplest possible case, where a single trait directly serves as the proxy for another. GWAS were conducted among 350K unrelated subjects in the UK Biobank, and results were clumped to identify independent ($R^2 \leq 0.1$) genome-wide significant (GWS; $P < 5 \times 10^{-8}$) loci. Loci for the target trait are considered true positives while loci for the proxy trait are considered predicted positives. Overlap analysis was performed to determine how many loci for the target trait were overlapped by a locus for the proxy trait, and vice versa, from which the empirical TPR and FDR were calculated.

In the case of urate and gout, there were 956 loci for urate and 47 loci for gout. The small number of loci for the latter is attributable to gout being binary rather than continuous, and having low prevalence in our cohort (7.5K cases, 1.5% of the cohort). Viewing urate as the proxy phenotype and gout as the true phenotype, all 47 loci for gout were overlapped by a locus for urate, giving an empirical TPR of 100% (47/47 loci). Of the 956 loci for urate, 590 (61.7%) did not overlap with a locus for gout. There are two possible explanations for the appearance of nominal false positives in the setting of an upstream proxy phenotype. If the measured circulating urate phenotype is truly causal for gout, then the 590 loci detected for urate but not gout are in fact causal for gout, and the latter GWAS simply lacked power to detect them. Alternatively, the measured circulating urate phenotype may be insufficiently specific, for example because only the concentration of urate in the synovial fluid contributes to gout pathogenesis. Causally, this would correspond to a diagram like that in **Figure S1**, where B_1 (e.g. synovial urate) is the biomarker directly on the causal path for gout, and B_2 (e.g. circulating urate) is a genetically correlated biomarker not on the causal path. Studying a proxy phenotype based on B_2 would still allow for detection of true positives G_{12} affecting both B_1 and B_2, but will miss true positives G_1 affecting B_1 and not B_2 while introducing the possibility of detecting false positive variants G_2 that affect B_2 and not B_1.

In the case of height and weight, there were 4100 loci for height and 1118 loci for weight. To complement the analysis of urate and gout, we consider weight as a downstream proxy for height. Among the 4100 loci for height, 2308 were overlapped by a locus for weight, for an empirical TPR of 56.3%. Meanwhile, among the 1118 loci for weight, 321 did not overlap a locus for height, for an empirical FDR of 28.7%. Since any locus that affects height should have an effect on weight, failure to detect some height loci via weight is likely due to lack of power, perhaps because more of the variation in weight is attributable to the environment. Consistent with this hypothesis, the heritability of height, as estimated by LD score regression applied to our summary statistics, was 41.5%, compared with only 24.2% for weight. Conversely, those loci for weight that do not overlap a locus for height may affect weight via a pathway other than height, for instance by changing body composition.

3.2 Recovery of true positive variants depends on target phenotype heritability and proxy phenotype correlation

We next examine how the TPR depends on the quality of the proxy phenotype, as quantified by its phenotypic correlation with the true phenotype. For the simulation study, we focus on a data generating process in **Figure 3A**, where the proxy phenotype P^* is imputed from a biomarker B purely downstream of the true phenotype P. Other data generating processes lead to qualitatively similar conclusions. In the simulation, there are two sources of environmental variation, E_1 affecting P, and E_2 affecting B. We examine how the TPR depends on the heritability of P and the correlation between P and P^*. This correlation is in turn determined by the relative influence of P versus E_2 on the biomarker that enters the imputation model. For simplicity, we let B itself act as the proxy phenotype. The results, presented in **Figure 3B**, demonstrate that success in recovering the causal variants increases with the heritability of the true phenotype and with the correlation of P with P^*. Intuitively, the TPR increases as the variation in the proxy phenotype explained by the variants G causal for the true phenotype P increases. The TPR decreases as more of the variation in P^* is explained by the environment, either due to having a noisier phenotype (E_1) or due to having a noisier biomarker (E_2).

To illustrate these trends with real data, we selected multiple phenotypes from [25] with heritabilities ranging from 5% for chronological age to 41% for mean platelet volume. Downstream proxy phenotypes having specified correlation with the true phenotype were generated by adding mean-zero noise to the true phenotype, via equation (7), such that the correlation between P and P^* was set to $\rho \in (0.05, 0.10, \ldots, 0.95)$. The TPR was measured as the proportion of phenotype genome-wide significant for the original P recovered via GWAS of P^*. The results, depicted in **Figure 3C**, recapitulate the trends from the simulation study. Recovery of the variants causal for the original phenotype increased with the heritability of P and with the correlation between P and P^*.

3.3 Why phenotypic correlation is a misleading indicator of utility for genetic discovery

The previous experiment suggests that power to detect genetic variants associated with the target phenotype P by means of a proxy phenotype P^* generally increases with the phenotypic correlation between P and P^*. However, as we show next, the source of this correlation matters. A proxy phenotype whose correlation with the target is purely environmental rather than genetic in origin will not assist in identifying genetic variants associated with P, regardless of the correlation between P and P^*. To illustrate this, we simulated a target phenotype P and a biomarker B according to the data generating process in **Figure 4A**. In contrast to our previous case studies, here the biomarker is neither directly upstream nor downstream of the target phenotype. The proportion of variants having causal effects on both P and B was varied between 0% and 100%, as was the magnitude of the environmental correlation. The heritabilities of both P and B were fixed at 50%, and the genetic correlation of P and B across the subset of variants causal for both phenotypes was fixed at $\rho = 0.5$. The biomarker served as the proxy phenotype.

Figure 4B illustrates how the phenotypic correlation between P and P^* increases with both the proportion of shared causal variants and with the environmental correlation. **Figures 4C** and **4D** demonstrate that increasing the phenotypic correlation by increasing the proportion of shared causal variants increases the TPR and decreases the FDR. Meanwhile, increasing the phenotypic correlation by increasing the environmental correlation has no impact on either the TPR or FDR. Taken together, these results imply that having high phenotypic correlation with the target phenotype is not sufficient for a proxy phenotype to be useful for genetic discovery. In fact, having a high phenotype correlation is also not necessary. Proxy phenotypes with only modest phenotypic correlation but high genetic overlap (i.e. phenotypes in the upper left of **Figure 4B**) will provide greater TPRs and lower FDRs than phenotypes with high phenotypic correlations but poor genetic overlap (i.e. phenotypes in the lower right of **Figure 4B**).

To emphasize the distinction between genetic and environmental correlation, we simulated a phenotype P_G that was predominantly genetic in origin (99% heritability) and a phenotype P_E that was entirely environmental in origin (0% heritability). The data generating processes are shown in **Figure 5A**. A biomarker for P_G was created by adding noise to the genetic component of P_G, and a biomarker for P_E was created by adding noise to the environmental component of P_E. **Figure 5B** demonstrates that the genetic correlation of the P_G^* with P_G and of P_E^* with P_E is stable across a broad range of phenotypic correlations. This means that phenotypic correlation is not an effective substitute for genetic correlation. To illustrate this point with real world data, we constructed a proxy for a highly heritable phenotype by adding noise to a polygenic score (PGS) of height, and a proxy for a minimally heritable phenotype by adding noise to creatinine levels imputed from age, genetic sex, and estimated alcohol intake; creatinine levels are known to be strongly influenced by alcohol intake [26, 27]. **Figure 5C** demonstrates that, as in the simulation study, phenotypic correlation can be toggled independently of genetic correlation. Consequently, we argue that 'test set R^2', a common metric for evaluating machine-learning predictions in traditional settings [28], is a poor measure of imputation quality in the context of genetic discovery because it fails to disentangle the genetic signal from the environmental noise.

4 Conclusions

Machine-learning (ML) based imputation presents both opportunities and challenges for genetic discovery. The ability to accurately impute difficult-to-ascertain phenotypes from available surrogate data enables researchers to fill in missing values and augment data sets with unmeasured phenotypes of interest. However, care is needed to ensure the validity of the inferred genetic relationships when conducting genetic association studies with the outputs of ML-models [13, 14, 15]. Here we considered the increasingly common practice of performing GWAS on machine-learned proxy phenotypes. Our analyses of simulated and real data illustrate that considering the causal relationships among genotypes, the phenotype of interest, and any biomarkers input to an ML model is essential to understanding the operating characteristics of proxy GWAS. For example, when the imputation is based on biomarkers

known to lie on the causal upstream path of the target phenotype, we can be confident that any associations detected will be relevant to the phenotype of interest. Conversely, when the biomarkers are causally downstream of the target phenotype, proxy GWAS will retain power for detecting true positives but will likely incur contamination by false discoveries. Our work suggests that selecting the appropriate biomarkers for imputation is non-trivial and highly consequential. While using upstream biomarkers for imputation tends to offer better control over false discovery rates, downstream biomarkers can potentially provide higher power for detection, albeit with increased risk of false positives. The strategic choice between these options requires careful consideration of the trade-offs involved and expert knowledge of the underlying pathophysiology of disease. While the true causal relationships between biomarkers and phenotypes may not be always known, large scale mendelian randomization analysis across all biomarkers can provide a rough approximation of what biomarkers might be upstream and which ones might be downstream.

We also demonstrated that although having a higher correlation between the target and proxy phenotypes is generally desirable, the source of the correlation is more important than its magnitude. A proxy phenotype with lower absolute phenotypic correlation that is driven predominantly by genetic overlap will provide greater utility for target-phenotype genetic discovery than a proxy phenotype with higher absolute phenotypic correlation that is driven predominantly by environmental factors. This finding cautions against use of the test set R^2 as a solitary measure of phenotype quality due to its inability to differentiate between genetic and environmental correlations. When evaluating machine-learned proxy phenotypes, having high genetic correlation with the target phenotype is the most desirable scenario, as it suggests the target and proxy phenotypes share many associated variants in common, and that the estimated directions of effect are consistent. Conversely, having a high phenotypic correlation but low genetic correlation is undesirable, as it suggests the correlation is driven by environmental factors, and that many variants associated with the proxy phenotype may not be relevant to the target phenotype. Global genetic correlation, however, is an imperfect indicator of whether a candidate proxy phenotype has high genetic overlap with the target phenotype, as shown by recent work on retinal epithelium pigmentation versus thickness [29]. For the purposes of genetic discovery, the ideal proxy phenotype will associate with all variants that affect the target phenotype, and no variants that do not affect the target phenotype, but need not have the same magnitude or direction of effect. Adding the latter requirements (i.e. same magnitude and direction of effect) would imply a strong (global) genetic correlation with the target and proxy phenotypes.

While this paper has focused on applications of ML in imputing unmeasured or missing target phenotypes, another emerging use of ML in GWAS is to derive lower dimensional representations of high dimensional phenotypes [12, 30, 31, 32, 33, 34]. Due to the difference in the problem specification, careful consideration is likely needed to define what constitute true and false positives for such phenotypes. An important direction of future work is to explore the causal considerations that determine the utility of representation phenotypes for genetic discovery.

Acknowledgements

The authors would like thank the participants of the UK Biobank, whose data were used with permission. This research was conducted using the UK Biobank Resource under approved Application Number 51766.

References

[1] Abdel Abdellaoui, Loic Yengo, Karin JH Verweij, and Peter M Visscher. 15 years of gwas discovery: realizing the promise. *The American Journal of Human Genetics*, 110(2):179–194, 2023.

[2] All of Us Research Program Genomics Investigators. Genomic data in the all of us research program. *Nature*, 627(8003):340–346, 2024.

[3] Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, 50(3):390–400, 2018.

[4] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599(7886):628–634, 2021.

[5] Mitja I Kurki, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, Hannele Laivuori, Mervi Aavikko, Mari A Kaunisto, et al. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, 2023.

[6] Wei Zhou, Masahiro Kanai, Kuan-Han H Wu, Humaira Rasheed, Kristin Tsuo, Jibril B Hirbo, Ying Wang, Arjun Bhattacharya, Huiling Zhao, Shinichi Namba, et al. Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10), 2022.

[7] Mary E Haas, James P Pirruccello, Samuel N Friedman, Minxian Wang, Connor A Emdin, Veeral H Ajmera, Tracey G Simon, Julian R Homburger, Xiuqing Guo, Matthew Budoff, et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell genomics*, 1(3), 2021.

[8] Babak Alipanahi, Farhad Hormozdiari, Babak Behsaz, Justin Cosentino, Zachary R McCaw, Emanuel Schorsch, D Sculley, Elizabeth H Dorfman, Paul J Foster, Lily H Peng, et al. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *The American Journal of Human Genetics*, 108(7):1217–1230, 2021.

[9] James P Pirruccello, Mark D Chaffin, Elizabeth L Chou, Stephen J Fleming, Honghuang Lin, Mahan Nekoui, Shaan Khurshid, Samuel F Friedman, Alexander G Bick, Alessandro Arduini, et al. Deep learning enables genetic analysis of the human thoracic aorta. *Nature genetics*, 54(1):40–51, 2022.

[10] Andrew Dahl, Michael Thompson, Ulzee An, Morten Krebs, Vivek Appadurai, Richard Border, Silviu-Alin Bacanu, Thomas Werge, Jonathan Flint, Andrew J Schork, et al. Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nature Genetics*, 55(12):2082–2093, 2023.

[11] Hari Somineni, Sumit Mukherjee, David Amar, Jingwen Pei, Karl Guo, David Light, Kaitlin Flynn, insitro Research Team, Chris Probert, Thomas Soare, et al. Machine learning across multiple imaging and biomarker modalities in the uk biobank improves genetic discovery for liver fat accumulation. *medRxiv*, pages 2024–01, 2024.

[12] Sumit Mukherjee, Zachary R McCaw, Jingwen Pei, Anna Merkoulovitch, Tom Soare, Raghav Tandon, David Amar, Hari Somineni, Christoph Klein, Santhosh Satapati, et al. Embedgem: A framework to evaluate the utility of embeddings for genetic discovery. *Bioinformatics Advances*, 4(1):vbae135, 2024.

[13] Zachary R McCaw, Jianhui Gao, Xihong Lin, and Jessica Gronsbell. Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nature Genetics*, pages 1–10, 2024.

[14] Jiacheng Miao, Yixuan Wu, Zhongxuan Sun, Xinran Miao, Tianyuan Lu, Jiwei Zhao, and Qiongshi Lu. Valid inference for machine learning-assisted gwas. *medRxiv*, pages 2024–01, 2024.

[15] Jessica Gronsbell, Jianhui Gao, Yaqi Shi, Zachary R. McCaw, and David Cheng. Another look at inference after prediction, 2024.

[16] Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Post-prediction inference. *BioRxiv*, pages 2020–01, 2020.

[17] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[18] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[19] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[20] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[21] B Bulik-Sullivan, HK Finucane, V Anttila, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236 – 1241, 2015.

[22] ZR McCaw, JM Lane, R Saxena, S Redline, and X Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262 – 1272, 2020.

[23] M Cadzow, TR Merriman, and N Dalbeth. Performance of gout definitions for genetic epidemiological studies: analysis of uk biobank. *Arthritis Res Ther*, 19:181, 2017.

[24] Remi Daviet, Gökhan Aydogan, Kanchana Jagannathan, Nathaniel Spilka, Philipp D Koellinger, Henry R Kranzler, Gideon Nave, and Reagan R Wetherill. Associations between alcohol consumption and gray and white matter volumes in the uk biobank. *Nature communications*, 13(1):1175, 2022.

[25] Jie Zheng, A Mesut Erzurumluoglu, Benjamin L Elsworth, John P Kemp, Laurence Howe, Philip C Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Early Genetics, Lifecourse Epidemiology (EAGLE) Eczema Consortium, et al. Ld hub: a centralized database and web interface to perform ld score regression that maximizes the potential of summary level gwas data for snp heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279, 2017.

[26] Yahia Sayed Mohamed, Suhair A Ahmed, Siddig B Mohamed, and AbdElkarim A Abdrabo. Evaluation of alcoholic consumption on serum uric acid, urea, and creatinine levels. *Ejpmr*, 3(5):577–579, 2016.

[27] Sisir K Majumdar, GK Shaw, P O'Gorman, and Allan D Thomson. Plasma urea and creatinine status in chronic alcoholics. *Drug and Alcohol Dependence*, 9(2):97–100, 1982.

[28] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[29] T Julian, T Fitzgerald, UK Biobank Eye, Vision Consortium, E Birney, and PI Sergouniotis. Pigmentation and retinal pigment epithelium thickness: a study of the phenotypic and genotypic relationships between ocular and extraocular pigmented tissues. *bioRxiv*, 2024.

[30] Sumit Mukherjee, Laura Heath, Christoph Preuss, Suman Jayadev, Gwenn A Garden, Anna K Greenwood, Solveig K Sieberts, Philip L De Jager, Nilüfer Ertekin-Taner, Gregory W Carter, et al. Molecular estimation of neurodegeneration pseudotime in older brains. *Nature communications*, 11(1):5781, 2020.

[31] Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, et al. Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. *Nature Genetics*, 56(8):1604–1613, 2024.

[32] J Cosentino, B Behsaz, B Alipanahi, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nature Genetics*, 55(5):787 – 795, 2023.

[33] Shubham Chaudhary, Almut Voigts, Michael Bereket, Matthew L Albert, Kristina Schwamborn, Eleftheria Zeggini, and Francesco Paolo Casale. Histogwas: An ai-enabled framework for automated genetic analysis of tissue phenotypes in histology cohorts. *bioRxiv*, pages 2024–06, 2024.

[34] Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, et al. igwas: image-based genome-wide association of self-supervised deep phenotyping of human medical images. *medRxiv*, pages 2022–05, 2022.

Figures

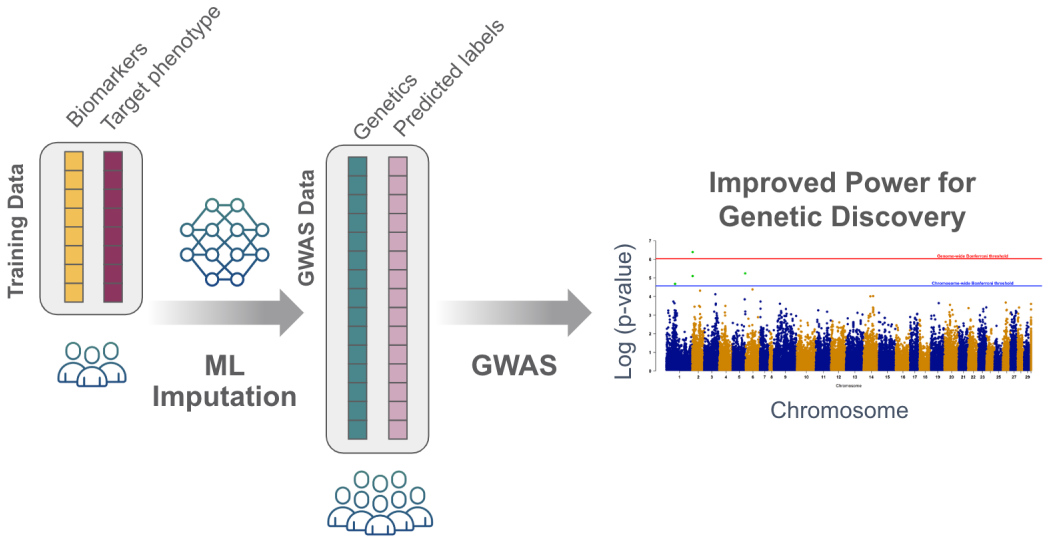


Figure 1: **Typical setup for machine learning (ML)-assisted proxy GWAS**. Within a labeled training data set, surrogate biometric data (e.g. biomarkers) is utilized to develop models for imputing the target phenotype. Once trained, the model extrapolates phenotypic labels across a larger dataset, augmenting the effective sample size for genome-wide association studies (GWAS). Since the predicted labels are used in place of measured, ground-truth labels, this strategy is referred to as proxy GWAS. The goal is to leverage the imputed proxy phenotype to identify genetic variants associated with the original target phenotype.

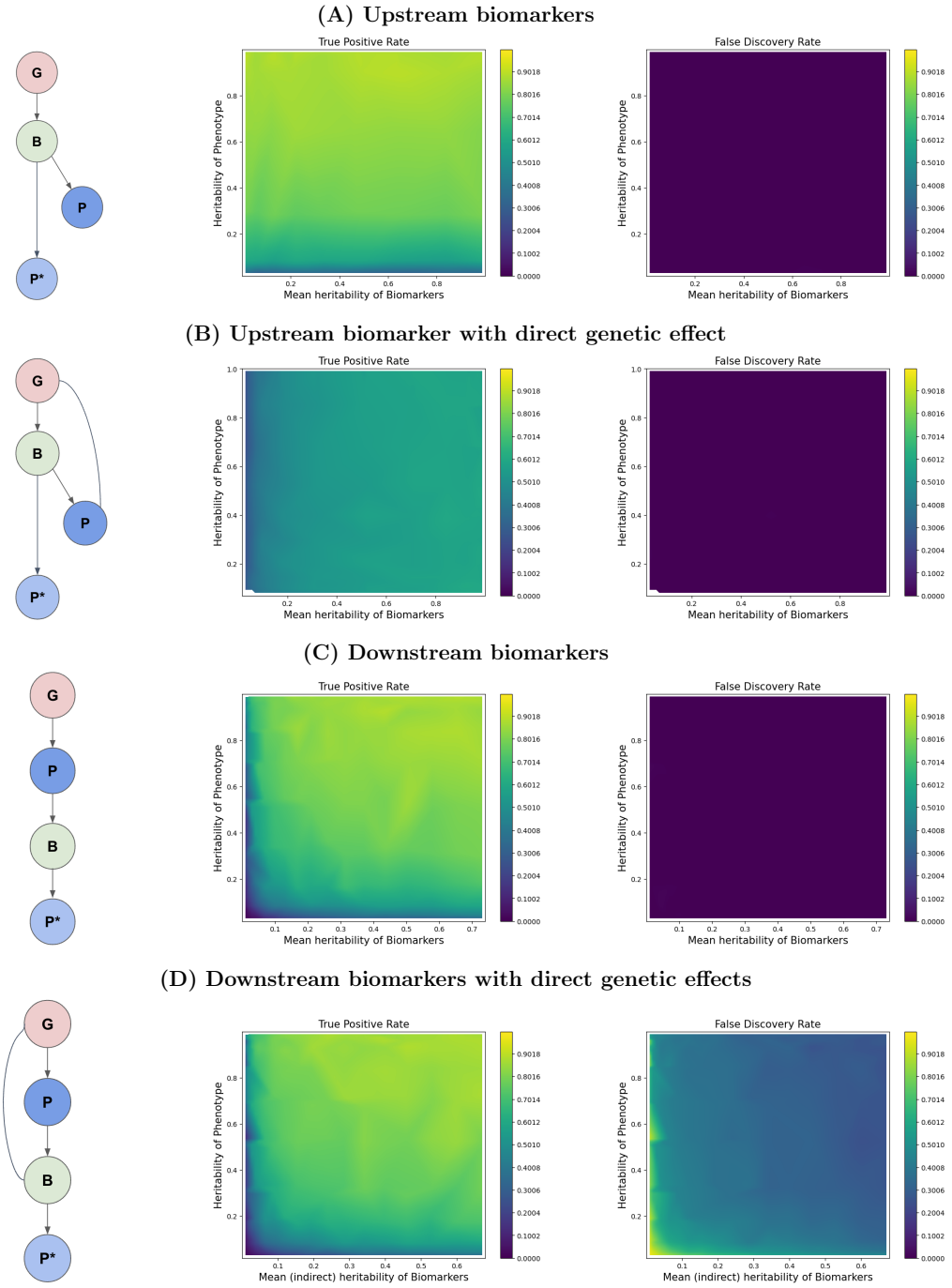


Figure 2: **The operating characteristics of proxy GWAS depend on the causal relationship between the true and proxy phenotypes.** Four different causal scenarios are show. In (A) and (B), the biomarkers B causally upstream of the true phenotype P, while in (C) and (D) the biomarkers are causally downstream. In all cases, the proxy phenotype P^* is derived from the biomarkers. The edges in the causal diagrams on the left show direct effects. The true positive rate in the central column is the proportion of variants causal for P that are detected by GWAS for P^*. The false discovery rate in the right column is the proportion of genome-wide significant variants for P^* that are not causal for P.

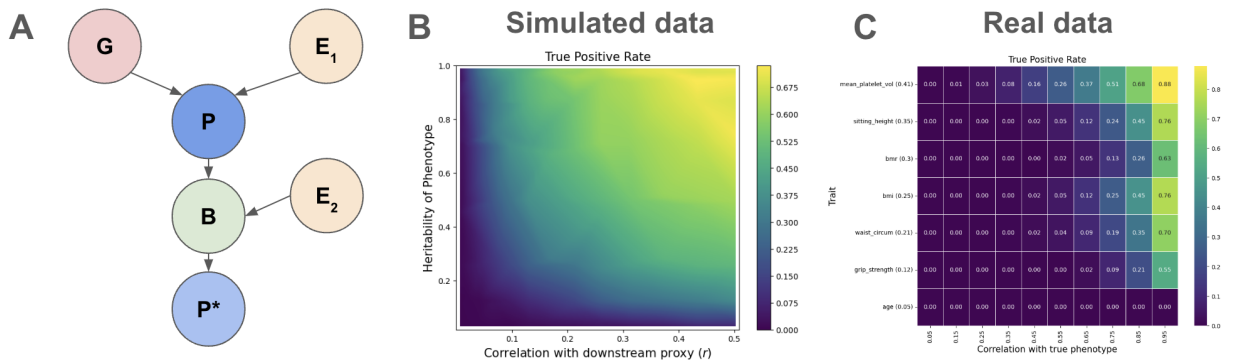


Figure 3: **True positive rate increases with target phenotype heritability and proxy phenotype correlation**. (A) Causal diagram for the data generating process used in the simulation study. Here G is genotype, P is the true phenotype, which is affected by environmental factor E_1, B is the biomarker, which is affected by environmental factor E_2, and P^* is the proxy phenotype. (B-C) True positive rate as the proportion of causal variants for P recovered as a function of the heritability of P and the phenotypic correlation between P and P^*. (B) presents results on simulated data, and (C) on real data, where mean-zero noise was introduced to control the correlation between the true and proxy phenotypes.

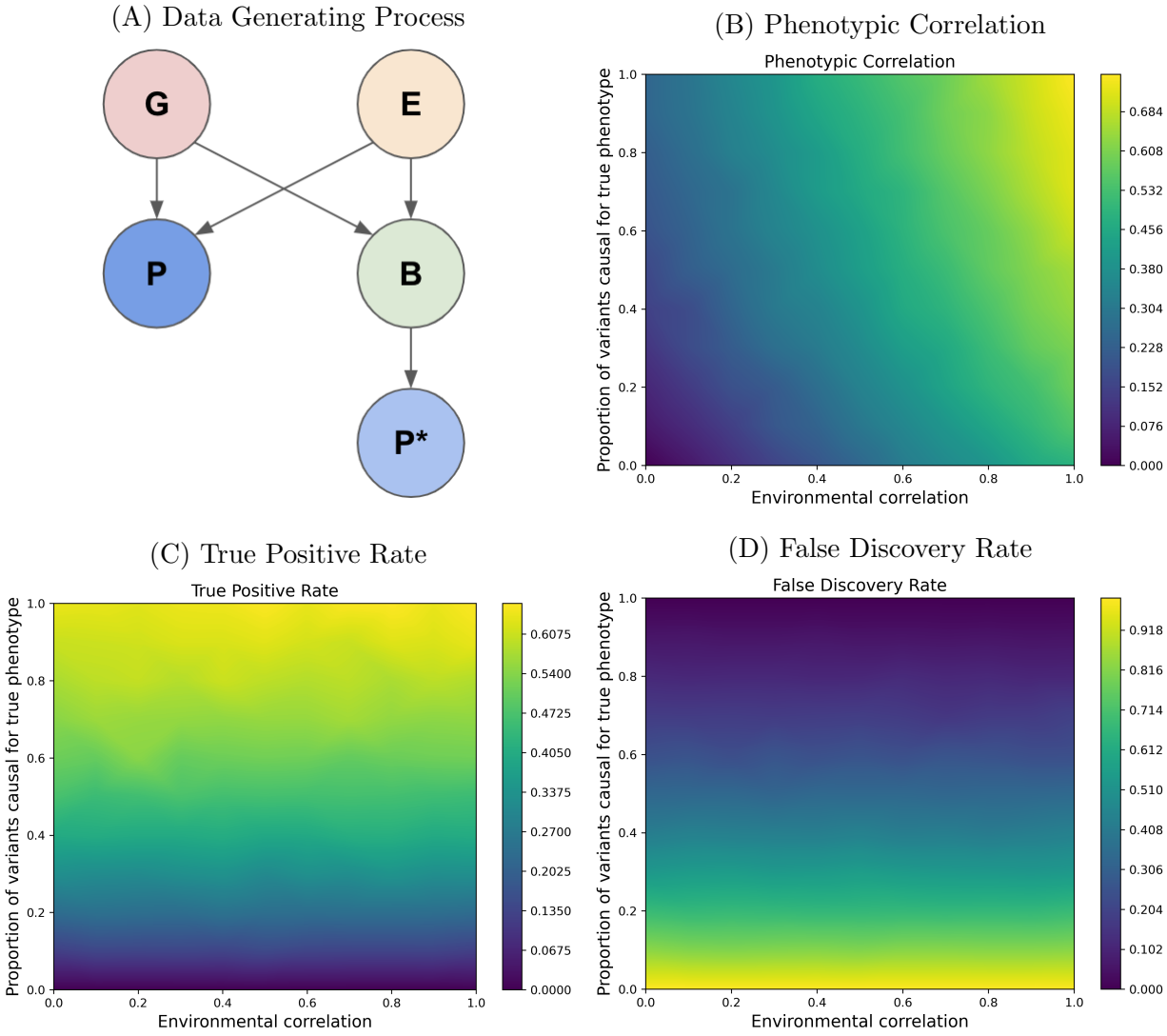
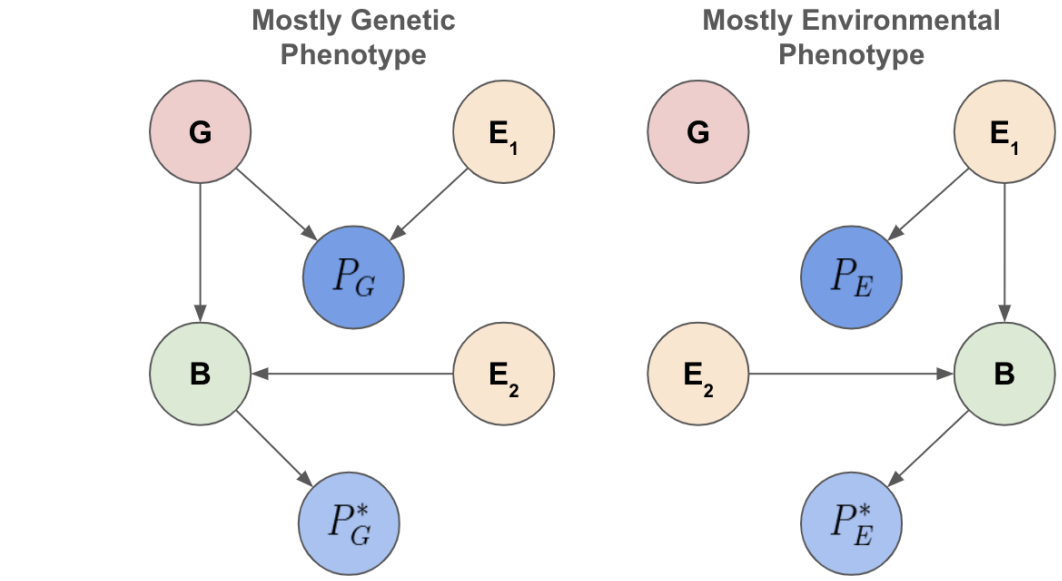
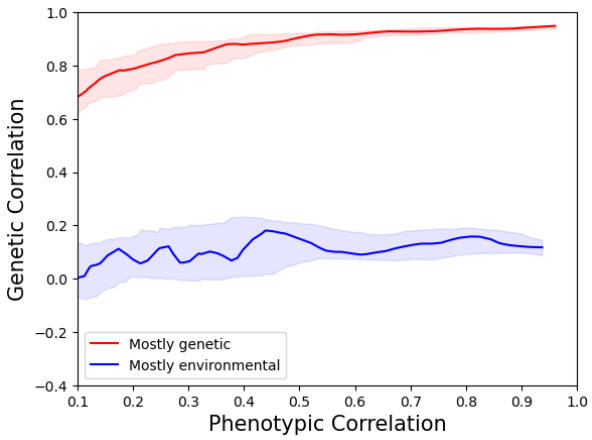


Figure 4: **Common genetic basis rather than raw phenotypic correlation determines utility for genetic discovery.** (A) Causal diagram for the data generating process. Here G is genotype, E is the environment, P is the target phenotype, B is the biomarker, and P^* is the proxy phenotype. Both P and B have a heritability of $h^2 = 50\%$ and a genetic correlation of zero. The proportion of variants affecting both P and B and the environmental correlation were each varied between 0 and 100%. (B) Phenotypic correlation as a function of the proportion of shared variants and the environmental correlation. (C) True positive rate and (D) false discovery rate as a function of the proportion of shared variants and the environmental correlation.

(A) Data Generating Process



(B) Simulated Data



(C) Real Data

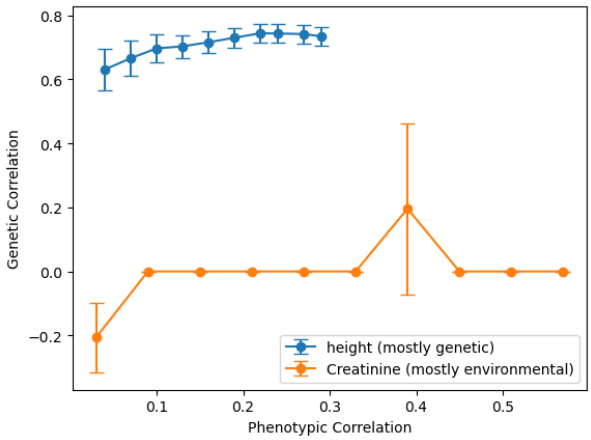


Figure 5: **Phenotypic correlation can vary independently of genetic correlation.** (A) Causal diagram for the data generating process. Here G is genotype, P_G and P_E represent a phenotype that is predominantly genetic or environmental in origin, respectively, P_G^* and P_E^* are corresponding proxy phenotypes, B is a biomarker, and E is environmental noise. (B) and (C) show genetic correlation versus phenotypic correlation for simulated (B) and real (C) data. For the real data setting, a proxy for a highly heritable phenotype was created by adding noise to a polygenic score for height, and a proxy for a minimally heritable phenotype was created by adding noise to circulating createnine imputed from age, genetic sex, and estimated alcohol intake.