A Bayesian method for estimating gene-level polygenicity under the framework of transcriptome-wide association study

Arunabha Majumdar¹ | Bogdan Pasaniuc²

¹Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Telangana, India
²Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, California

Correspondence
Arunabha Majumdar, Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Telangana, India.
Email: arun.majum@math.iith.ac.in

Funding information
UK Biobank, Grant/Award Numbers: 33297, 33127

Polygenicity refers to the phenomenon that multiple genetic variants have a nonzero effect on a complex trait. It is defined as the proportion of genetic variants with a nonzero effect on the trait. Evaluation of polygenicity can provide valuable insights into the genetic architecture of the trait. Several recent works have attempted to estimate polygenicity at the single nucleotide polymorphism level. However, evaluating polygenicity at the gene level can be biologically more meaningful. We propose the notion of gene-level polygenicity, defined as the proportion of genes having a nonzero effect on the trait under the framework of a transcriptome-wide association study. We introduce a Bayesian approach genepoly to estimate this quantity for a trait. The method is based on spike and slab prior and simultaneously estimates the subset of non-null genes. Our simulation study shows that genepoly efficiently estimates gene-level polygenicity. The method produces a downward bias for small choices of trait heritability due to a non-null gene, which diminishes rapidly with an increase in the genome-wide association study (GWAS) sample size. While identifying the subset of non-null genes, genepoly offers a high level of specificity and an overall good level of sensitivity—the sensitivity increases as the sample size of the reference panel expression and GWAS data increase. We applied the method to seven phenotypes in the UK Biobank, integrating expression data. We find height to be the most polygenic and asthma to be the least polygenic.

KEYWORDS
complex trait genomics, gene-level association, hierarchical models, MCMC, spike and slab prior

1 | INTRODUCTION

Beyond the discovery of genetic loci associated with a complex trait, it is crucial to estimate the overall distribution of the effect size of genome-wide genetic variants to understand the genetic architecture of the trait better.¹ Polygenicity is the proportion of genetic variants with a nonzero effect on the trait. It is an essential characteristic of the effect size distribution and provides valuable insights into the genetic architecture of the trait. Efficient estimation of this quantity can help to improve the design of risk prediction models² or reveal the biological complexity of a trait.³ Several approaches have attempted to estimate the polygenicity using its definition as the proportion of single nucleotide polymorphisms (SNPs), which have a nonzero effect on the trait under the framework of genome-wide association study (GWAS).¹,⁴ We refer to
it as SNP-level polygenicity. However, defining polygenicity at a gene level can be biologically more meaningful. In this article, we propose the notion of gene-level polygenicity, which we define as the proportion of genes that have a nonzero effect on the trait. An expression data set on the trait of interest would be ideal for estimating the gene-level polygenicity. However, most expression data sets have a limited sample size, and the data on various quantitative or disease traits of interest may not be available.

GWAS has successfully associated SNP loci with many complex traits. However, most of the statistically significant GWAS signals are located in the non-coding regions of the human genome. It is challenging to prioritize the causal genes that mediate the effects of these non-coding variants on the trait from GWAS data. Non-coding regions are likely to have regulatory effects on gene expression. Expression quantitative trait loci (eQTLs) are an important class of regulatory variants. Common variants associated with a complex trait tend to co-localize with the eQTLs of the genes that are causal for the trait. Transcriptome-wide association study (TWAS) aggregates regulatory effects of multiple eQTLs and tests for association between a gene and a trait.

For example, consider the lipid trait, LDL, the complex trait of interest, and liver tissue as the relevant tissue type. The GTEx data has measurements of gene expressions in various tissue types and genotypes for genome-wide SNPs. In the first step of TWAS, consider a gene and its expression in the liver tissue in the GTEx data. Next, consider a set of SNPs surrounding the gene (local SNPs), for example, within 500 kilobases or one megabase region from the gene boundaries. We fit a regression model of expression on local SNPs’ genotypes to learn a prediction model for expression. In the second step of TWAS, we have GWAS data for a different set of individuals. For example, UK Biobank has data for LDL and genotypes for genome-wide SNPs, but expression data is yet to be available. We implement the prediction model of expression obtained in the GTEx data to impute the expression for the GWAS individuals in the UK Biobank. Finally, we regress LDL on the predicted genetically regulated expression component to test for an association between LDL and the gene. In the absence of individual-level LDL and genotype data for the GWAS individuals, TWAS versions can be implemented using the summary-level GWAS association data, the prediction model of expression, and a publicly available genotype database, such as 1000 genome data to estimate the LD structure of SNPs. TWAS has few merits compared to GWAS. First, the gene-level association is biologically more interpretable than a single SNP-level association. Second, gene-level tests have a lower multiple testing burden due to a much smaller number of genes (in the order of thousands) than the number of SNPs (in the order of hundreds of thousands). Third, TWAS is a locus-level test that integrates multiple variants to test for an association of a gene. A recent work has demonstrated that the relative success of TWAS compared to GWAS can vary across genes. If the expression of a gene is more heritable, TWAS can be more effective than GWAS. However, if the expression heritability is low, GWAS can be more effective than TWAS.

Considering the above deliberations, we develop a statistical approach to estimate the gene-level polygenicity under the framework of TWAS using summary-level association data. While evaluating SNP-level polygenicity, a crucial step is to account for the linkage disequilibrium (LD) among SNPs and identify the SNPs which originally have a nonzero effect. Two current methods, Genesis and BEAVR, estimate the number/proportion of susceptibility SNPs for a trait while considering LD structure. Genesis forecasts the number of nonzero effect SNPs at a genome-wide level, which can be converted into a proportion of non-null SNPs. BEAVR partitions the genome into regions and estimates regional polygenicity. A marginally associated SNP may not always have a nonzero effect. Instead, it may be in LD with an SNP, which has a nonzero effect. Similarly, a marginally associated gene in TWAS may not originally have a nonzero effect. Instead, the marginal association may be due to a correlation between the predicted expression of the gene and another gene (gene co-regulation) where the latter truly has a nonzero effect. Such correlation between two genes’ predicted expression can arise due to shared eQTLs or LD between eQTLs of the genes. We develop a Bayesian approach to estimate the proportion of genome-wide genes, the genetic component of expressions of which have a nonzero effect on the trait (non-null genes). We consider this proportion of non-null genes as a measure of gene-level polygenicity. We estimate the proportion of non-null genes explicitly accounting for the correlation between the predicted genetic components of expressions. Our unified Bayesian framework also simultaneously estimates the true subset of non-null genes. We refer to the method as genepoly (gene-level polygenicity). We consider a continuous spike and slab prior to developing genepoly. The spike component represents the null effect, and the slab component represents the non-null effect. We perform a fully Bayesian inference based on MCMC while explicitly accounting for the covariance structure among the genes. Furthermore, the method uses marginal TWAS summary statistics, which are often publicly available. As an example, to apply our approach for LDL, previously discussed for TWAS, the key ingredients are the same as that needed for a TWAS based on summary statistics, that is, summary-level GWAS association data, expression prediction models, and a publicly available genotype database, for example, 1000 genome data to estimate the LD structure of SNPs.
We perform extensive simulations to evaluate the performance of the approach. *genepoly* efficiently estimates the gene-level polygenicity under various simulation scenarios. It produces a downward bias in the estimation when the heritability per non-null gene is small. The downward bias decreases rapidly as the GWAS sample size increases. Simulations also show that the q-value approach implemented in this context consistently produces a sizeable upward bias in all simulation scenarios. While identifying the subset of non-null genes, *genepoly* offers high specificity and good sensitivity across the simulation scenarios. The sensitivity improves with the increased sample size of the reference panel expression and GWAS data. We applied *genepoly* to seven traits in the UK Biobank integrating expression data. Our analysis shows that height is the most polygenic, and asthma is the least polygenic. Our analyses also suggest that HDL and triglycerides are more polygenic than LDL.

## 2 MATERIALS AND METHODS

### 2.1 Overview of methods

The standard TWAS consists of two-stage regressions. In the first stage, we consider a tissue of interest in the reference panel of expression (and genotype) data. We fit a penalized regression\(^{18,19}\) to evaluate the effect of genotypes of SNPs surrounding the gene (local SNPs) on the expression. From the regression for each gene, we obtain a prediction model to estimate the genetic component of the gene’s expression based on its local SNPs. In the second stage, we use the prediction model to predict the genetic component of expression in the GWAS data based on the same set of local SNPs. We then regress a GWAS trait on the predicted expression to assess an association between the gene and the trait. We repeat this pipeline to obtain the marginal TWAS statistics for all the genes. Next, we derive the analytic formulas of the expectation vector and covariance matrix for the TWAS statistics of all genes. We assume a multivariate normal distribution of the TWAS statistics for the genes on each chromosome. Our main goal is to estimate the proportion of all such genome-wide genes with a nonzero effect on the trait (gene-level polygenicity).

We also aim to identify the true subset of non-null genes. To model sparsity, we consider a continuous spike and slab prior distribution\(^{14,17}\) for the TWAS effect sizes and develop a unified Bayesian approach to perform both the estimation of gene-level polygenicity and the selection of non-null genes. With prior probability \(p\), the TWAS effect size follows the slab distribution representing a non-null effect. In the data likelihood, we explicitly account for the covariance structure of genes. We derive the full-conditional posterior distributions of the model parameters to implement MCMC using Gibbs sampling. Finally, we perform Bayesian inference based on the posterior sample of the model parameters obtained by the MCMC. Next, we define some notations that we use while describing the methods section.

#### 2.1.1 Various notations used for the reference panel expression data

- \(m\): Number of genes on a chromosome, \(n_E\): number of individuals.
- \(E_j\): Expression of \(j\)th gene in the specific tissue of interest, \(j = 1, \ldots, m\).
- \(E_{ij}\): Expression of \(j\)th gene for \(i\)th individual, \(i = 1, \ldots, n_E, j = 1, \ldots, m\).
- \(r_j\): Number of local SNPs for \(j\)th gene, \(j = 1, \ldots, m\).
- \(r\): total number of distinct local SNPs.
- \(x_{j,\text{ref}}\): Genotype vector for local SNPs of \(j\)th gene, \(j = 1, \ldots, m\).
- \(x_{ij,\text{ref}}\): Genotype vector for local SNPs of \(j\)th gene for \(i\)th individual, \(i = 1, \ldots, n_E, j = 1, \ldots, m\).
- \(w_j\): Effect size vector of local SNPs on the expression of \(j\)th gene, \(j = 1, \ldots, m\).

#### 2.1.2 Various notations used for the GWAS data

- \(n\): Number of individuals. In the simulation section, we replace \(n\) by \(n_{GW}\).
- \(Y\): Vector of trait values for the individuals in GWAS.
- \(X\): Genotype matrix of order \(n \times r\) for GWAS individuals at all local SNPs on a chromosome.
- \(G\): Genetic component of expressions for \(m\) genes in GWAS. \(G\) is \(n \times m\) ordered matrix.
- \(\alpha\): Joint effect size vector of the predicted genetic component of expressions for \(m\) genes in the GWAS data, \(\alpha = (a_1, \ldots, a_m)\). \(\lambda = \sqrt{n}\alpha\).
Regression of trait on predicted expression in GWAS data

Regression of gene expression on local SNPs in reference panel

We describe the main steps of genepoly for \( m \) genes on a single chromosome, the extension of which for all chromosomes is straightforward. In the reference panel data, we regress the expression of \( j \)th gene on the genotypes of its local SNPs (eg, SNPs within 0.5 MB window of the gene boundary), \( j = 1, \ldots, m \). We consider the following linear model:

\[
E_j = x_{j, \text{ref}}' w_j + \epsilon_j.
\]

\( E_j \) denotes the mean-centered expression of \( j \)th gene in the specific tissue of interest. For the local SNPs, \( x_j \) denotes the normalized genotype vector (centered for mean and then scaled by standard deviation), \( w_j \) denotes the effect size vector, and \( \epsilon_j \) denotes the random error. Since the number of local SNPs can be close to or larger than the sample size of the reference panel (eg, GTEx data), we implement a penalized regression (eg, Lasso, Elastic Net) to estimate \( w_j \) individually for each \( j = 1, \ldots, m \). Suppose we have \( r_j \) local SNPs for \( j \)th gene. Based on Lasso,

\[
\hat{w}_j = \arg \min_{w_j \in \mathbb{R}^{r_j}} \sum_{i=1}^{n_E} (E_{ij} - x_{i, \text{ref}}' w_j)^2 + \lambda_j \sum_{l=1}^{r_j} |w_{lj}|.
\]

Let \( r \) denote the total number of unique SNPs considered for the \( m \) genes, and the SNPs are arranged in increasing order of base pair positions. Consider \( r \times m \) matrix: \( \hat{W} = [\hat{w}_1, \ldots, \hat{w}_m] \). For \( j \)th gene, all entries of \( \hat{w}_j \) are zero except for the \( r_j \) local SNPs. Some of the \( r_j \) entries can be zero due to fitting a penalized regression (eg, Lasso). For each of these \( m \) genes, we also assume that the local SNPs produced a significantly positive heritability of expression. We term such a gene as a locally heritable gene. We drop a gene from the downstream analysis if it is not locally heritable.

Regression of trait on predicted expression in GWAS data

In general, expression measurements are not available in GWAS data. Suppose the genotype data for the set of local SNPs considered for a gene in the reference panel is also available in GWAS data. In that case, we can predict the genetic component of the gene’s expression using \( \hat{W} \) (obtained in the first stage regression). We estimate the genetic component of expression for \( m \) genes as \( \hat{G} = X \hat{W} \). \( X_{n \times r} \) is the genotype matrix for \( n \) individuals and \( r \) local SNPs in the GWAS data. The genotype data of each SNP is normalized to have zero mean and unity variance. Subsequently, we can perform a multiple linear regression of a continuous GWAS trait \( Y \) on \( \hat{G} \) to evaluate the joint effect of the \( m \) genes on \( Y \) as:

\[
Y = \hat{G} \alpha + \epsilon = a_1 \hat{G}_1 + \cdots + a_m \hat{G}_m + \epsilon
\]

\( Y \) denotes the trait vector for \( n \) individuals. \( \alpha = (a_1, \ldots, a_m)' \) denotes the joint effect sizes for the \( m \) genes, \( \hat{G}_j \) denotes the predicted expression for \( j \)th gene, \( j = 1, \ldots, m \). We ignore the intercept term considering \( Y \) to be mean-centered. We note that \( \hat{G} \) is an estimate of true \( G \) and involves some uncertainty (due to variability in \( \hat{W} \)) which is ignored in a standard TWAS. Here, \( \epsilon \) denotes the error term and we assume that \( \mathbb{E}(\epsilon) = 0, \text{cov}(\epsilon) = \sigma^2_e I_{n \times n} \). The joint ordinary least square (OLS) estimate of \( \alpha \) is given by: \( \hat{\alpha} = \hat{G} \hat{G}^{-1} \hat{G} Y \). We note that an OLS estimate of \( \alpha \) should be reliable because the number of locally heritable genes on a single chromosome (eg, 500) is expected to be much smaller than a contemporary GWAS data sample size (eg, 10 000).

In the standard TWAS, we consider a univariate regression: \( E(Y) = \hat{G}_j \beta_j \) with the error variance \( \sigma^2_e \), where \( \hat{G}_j \) denotes the predicted genetic component of \( j \)th gene’s expression (\( j \)th column of \( \hat{G} \)). We test for a marginal association, \( H_0 : \beta_j = 0 \)
Expectation of the vector of marginal TWAS statistics

We derive the expectation of marginal TWAS statistics considering the joint model of $Y$ in Equation (2): $E(Y) = \mathbf{G}\alpha$. Thus,

$$
E(z_j) = \frac{\hat{G}_j' E(Y)}{\sqrt{\hat{G}_j' \hat{G}_j}} = \frac{\hat{G}_j' \hat{G}_j}{\sqrt{\hat{G}_j' \hat{G}_j}}.
$$

Define the diagonal matrix $M_1 = \text{diag}(\frac{1}{\sqrt{G_{11}' G_{11}}}, \ldots, \frac{1}{\sqrt{G_{n}' G_{n}}})$. $\hat{G}' \hat{G} = \mathbf{W}' \mathbf{X}' \mathbf{W} = n\hat{W}'\hat{V}\hat{W}$. $V = \frac{1}{n} \mathbf{X}' \mathbf{X}$ is the LD matrix for the $r$ SNPs in GWAS data. Considering the vector of marginal $Z$ statistics across $m$ genes simultaneously, we obtain that:

$$
E(Z) = M_1 \hat{G}' \hat{G} \alpha = nM_1 \hat{W}'\hat{V}\hat{W} \alpha.
$$

Since $\hat{G}' \hat{G} = n\hat{W}'\hat{V}\hat{W}$, $j$th diagonal element of $\hat{G}' \hat{G}$ is given by: $\hat{G}_{j,j} = n[\hat{W}'\hat{V}\hat{W}]_{jj}$. We define a diagonal matrix, the diagonal entries of which are the diagonal elements of $\hat{W}'\hat{V}\hat{W}$: $\text{diag}(\hat{W}'\hat{V}\hat{W}) = \text{diag}([\hat{W}'\hat{V}\hat{W}]_{11}, \ldots, [\hat{W}'\hat{V}\hat{W}]_{mm})$. Thus, $M_1 = \frac{1}{\sqrt{n}} \{(\text{diag}(\hat{W}'\hat{V}\hat{W}))^{-1}\}$. Hence,

$$
E(Z) = \sqrt{\text{diag}(\hat{W}'\hat{V}\hat{W})^{-1}} \hat{W}'\hat{V}\hat{W} (\sqrt{n}\alpha).
$$

We denote $\lambda = \sqrt{n}\alpha$, $M_2 = \sqrt{\text{diag}(\hat{W}'\hat{V}\hat{W})^{-1}}$, and $S = \hat{W}'\hat{V}\hat{W}$. Thus,

$$
E(Z) = M_2 S \lambda.
$$

We note that $\alpha_j = 0$ is equivalent to $\lambda_j = 0, j = 1, \ldots, m$.

Covariance matrix of the vector of marginal TWAS statistics

We have $Z = \frac{1}{\sqrt{n}} M_2 \hat{G}' Y = \frac{1}{\sqrt{n}} M_2 \hat{G}' (\hat{G}\alpha + \hat{G}\epsilon)$. It can be derived that:

$$
\text{cov}(Z) = \sigma^2 M_2 S M_2, \text{ where } M_2 = \sqrt{\text{diag}(\hat{W}'\hat{V}\hat{W})^{-1}} \text{ and } S = \hat{W}'\hat{V}\hat{W}.
$$

We note that for a vector of $Z$ statistics, the correlation matrix is the same as the covariance matrix. The full covariance matrix for all chromosomes will be a block-diagonal matrix, where each block corresponds to each chromosome. The inverse of a block-diagonal matrix will again be a block-diagonal matrix containing the corresponding inverted sub-matrices. Two genes on a single chromosome residing far apart (e.g., >2 MB) will likely have a zero correlation. This induces a sliding-window type of covariance structure for the genes on each chromosome. We consider an inverse-gamma (IG) prior for $\sigma^2$ as $\sigma^2 \sim \text{IG}(d_1, d_2)$. In a later section, we discuss a sensible approach to choosing the hyper-parameters $d_1, d_2$. Next, we discuss the prior for the main parameter vector $\lambda$. 

2.6 Continuous spike and slab prior

Our main goal is simultaneously estimating the gene-level polygenicity and the true subset of non-null genes under a unified Bayesian framework. We consider a Bayesian approach based on continuous spike and slab prior,\cite{14,17} where the spike component represents a null effect, and the slab component represents a non-null effect. As mentioned above, we first describe the method for a set of \( m \) genes on a single chromosome which can easily be extended to all chromosomes.

From previous sections we have: \( E(Z) = M_2 S \lambda \) and \( \text{cov}(Z) = M_2 S M_2 \), where \( M_2 = \sqrt{\{\text{diag}(W'VW')\}^{-1}} \) and \( S = W'VW \). Note that the \( Z \) statistic for each gene follows normal. We assume that the vector of \( Z \) statistics follows a multivariate normal distribution.

\[
Z|\lambda \sim N_m(M_2S\lambda, \sigma^2 \Sigma M_2 S M_2).
\] (8)

We consider a continuous spike and slab prior distribution for \( \lambda \), the vector of TWAS effect sizes for \( m \) genes. For \( j = 1, \ldots, m \), we consider the prior of \( \lambda_j \) as:

\[
\lambda_j|c_j, \nu \sim (1-c_j)N(0, \sigma_0^2) + c_jN(0, \nu^2 \sigma_0^2); \ \nu \gg 1, \ \sigma_0^2 = 10^{-10}
\]

\[
P(c_j = 1|p) = p, \ P(c_j = 0|p) = 1 - p
\]

\[
p|a_1, a_2 \sim \text{Beta}(a_1, a_2)
\]

\[
\nu^2 \sigma_0^2 = \sigma^2 \sim \text{inverse gamma}(b_1, b_2).
\] (9)

The latent variable \( c_j \) defines whether the \( j \)th gene has a nonzero effect. When \( c_j = 0 \), \( \lambda_j \sim N(0, \sigma_0^2) \), and when \( c_j = 1 \), \( \lambda_j \sim N(0, \nu^2 \sigma_0^2) \). We consider a very small fixed value of \( \sigma_0^2 = 10^{-10} \) and a relatively much larger value of \( \sigma^2 = \nu^2 \sigma_0^2 \) (eg, 10) such that \( \nu \gg \frac{\sigma_0^2}{\sigma^2} \gg 1 \). Since \( \sigma_0^2 \) is nearly zero, if \( c_j = 0 \), \( \lambda_j \) would be very small and considered zero (\( j \)th gene is null). We can consider a sufficiently large value of \( \sigma^2 \) such that if \( c_j = 1 \), \( \lambda_j \) can be treated as nonzero (\( j \)th gene is non-null). We refer to \( \sigma_0^2 \) as the spike variance and \( \sigma^2 \) as the slab variance. A random variable following \( N(0, 10^{-10}) \) has 99% probability of taking a value between \((-2.6 \times 10^{-5}, 2.6 \times 10^{-5})\). The proportion of non-null genes, \( p \), measures the gene-level polygenicity. The vector of latent variables, \( C = (c_1, \ldots, c_m) \), defines the subset of non-null genes with a nonzero effect on the trait.

For convenience, we assume that \( \lambda_1, \ldots, \lambda_m \) are independently distributed in the prior. Conditioned on \( p \), the configuration indicators \( (c_1, \ldots, c_m) \) are i.i.d. Bernoulli(\( p \)), where \( p \sim \text{Beta}(a_1, a_2) \). We assume that \( \sigma^2 \sim \text{IG}(b_1, b_2) \). In a later section, we discuss a method of moments (MOM) approach to choose the hyperparameters \( (a_1, a_2) \) and \( (b_1, b_2) \). We perform full Bayesian inference on the gene-level polygenicity and the subset of non-null genes. Note that when \( \sigma_0^2 = 0 \), the prior has a positive mass at \( \lambda_j = 0 \) and is known as Dirac spike and slab prior.\cite{17}

We note that we can also develop the model starting with a multiple linear regression based on individual-level data and placing the spike and slab prior on the effect size vector (supplementary materials). This framework is known as the Bayesian variable selection regression.\cite{21} The supplementary materials outline the full-conditional posterior distribution of \( \lambda \) under this setup. We also discuss how we can modify this posterior distribution when we have summary-level GWAS association data only, and it turns out to be the same as the above full conditional posterior distribution of \( \lambda \).

2.7 Markov chain Monte Carlo

We implement the Markov chain Monte Carlo (MCMC) algorithm using Gibbs sampling to generate a posterior sample of the model parameters. We provide the full-conditional posterior distributions in the following. In a given MCMC iteration, we sequentially update \( \lambda, C, p, \sigma_1^2, \sigma_2^2 \).

**Full-conditional posterior distribution of \( \lambda \):** Define \( D_C \) to be a diagonal matrix with its \( j \)th diagonal element defined in the following way: if \( c_j = 0 \), it is chosen as \( \sigma_0^2 \), and if \( c_j = 1 \), it is chosen as \( \sigma^2 \). Thus, the diagonal entries of \( D_C \) are \( (\sigma_1^2, \ldots, \sigma_m^2) \). We obtain the full-conditional posterior distribution of \( \lambda \) as:

\[
\lambda|Z, C, p, \sigma_1^2, \sigma_2^2 \sim N_m(\mu, \Sigma), \text{ where } \Sigma = \left( \frac{1}{\sigma_0^2}S + D_C^{-1} \right)^{-1} \text{ and } \mu = \frac{1}{\sigma_0^2} \Sigma M_2^{-1} Z.
\]
Full-conditional posterior distribution of $C$: Let $C_{-j} = (c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_m)$. We update the configuration indicators using the full-conditional distribution of $c_j, j = 1, \ldots, m$.

$$P(c_j = 0| C_{-j}, Z, \lambda, p, \sigma^2_1, \sigma^2_2) = \frac{1}{1 + \frac{p}{1-p} f(\lambda_j|c_j = 1)}$$

Here, $f(\lambda_j|c_j = 1) = N(\lambda_j; 0, \sigma^2_1)$ and $f(\lambda_j|c_j = 0) = N(\lambda_j; 0, \sigma^2_0)$. $N(x; \mu, \sigma^2)$ denotes the normal density at $x$ given $\mu, \sigma^2$ to be the mean and variance. $P(c_j = 1| C_{-j}, Z, \lambda, p, \sigma^2_1, \sigma^2_2) = 1 - P(c_j = 0| C_{-j}, Z, \lambda, p, \sigma^2_1, \sigma^2_2)$.

Full-conditional posterior distribution of $p$: Denote $m_1 = \#\{c_j = 1, j = 1, \ldots, m\} = \sum_{j=1}^{m} c_j$ and $m_0 = m - m_1$. We update $p$ from: $p(Z, \lambda, C, \sigma^2_1, \sigma^2_2 \sim \text{Beta}(a_1 + m_1, a_2 + m_0)$.

Full-conditional posterior distribution of $\sigma^2_1$: Next, we update the slab variance $\sigma^2_1$ using the following IG distribution:

$$\sigma^2_1 | Z, \lambda, C, p, \sigma^2_2 \sim \text{IG}(\frac{m_1}{2}, \frac{m_1}{2} + \frac{1}{2} \sum_{j: c_j = 1} \lambda_j^2)$$

Note that, if $m_1 = 0$, $\sigma^2_1$ is updated from its prior IG($b_1, b_2$).

Full-conditional posterior distribution of $\sigma^2_2$: We update $\sigma^2_2$ from the following full conditional posterior distribution:

$$\sigma^2_2 | Z, \lambda, \sigma^2_1, C, p \sim \text{IG}(\frac{m_0}{2} + d_1, B + d_2),$$

where $B = \frac{1}{2}(Z - M_2 S \lambda)'(M_2 S M_2)^{-1}(Z - M_2 S \lambda)$.

2.8 Posterior inference

After a certain burn-in period of the MCMC, we collect the posterior sample of model parameters. We obtain various posterior summaries based on the MCMC sample. We use the posterior median of $p$ as the point estimate of the gene-level polygeneicity. We consider a 5%-95% central posterior interval of $p$.

2.9 Estimation of heritability

We outline an approach to estimating the trait’s heritability due to the genes’ predicted expressions. For an individual randomly selected from the sample, we let $y$ denote the phenotype value and $\hat{g}$ denote the predicted expression for the genes under consideration. Consider the linear model: $y' \hat{g} = \hat{g}' \alpha + e$, where $E(e) = 0$, $\text{var}(e) = \sigma^2_e$. Consider $\lambda = \sqrt{n} \alpha$ and $u = \frac{1}{\sqrt{n}} \hat{g}$. Hence, $y' u, \lambda = u' \lambda + e$, which implies that $E(y' u, \lambda) = u' \lambda$ and var($y' u, \lambda$) = var($e$) = $\sigma^2_e$.

Using the law of total variance, var($y' u$) = var($E(y' u, \lambda)$) + $E(\text{var}(y' u, \lambda))$ = var($u' \lambda$) + $\sigma^2_e$. We are interested in var($y$) and which proportion of it is explained by the systematic component. var($y$) = $E(\text{var}(y'u))$ + $E(\text{var}(y'u))$.

Note that, $E(y'u) = u'E(\lambda) + E(e) = 0$, since $E(\lambda) = 0$ in the prior. So, var($y$) = $E(\text{var}(y'u))$ = var($u' \lambda$) + $\sigma^2_e$. The heritability due to all the genes can be expressed as: $\sigma^2_H = \frac{\text{var}(u' \lambda)}{E(\text{var}(u' \lambda))}$. Again, applying the law of total variance conditioning on $u$, we obtain var($u' \lambda$) = $E(\text{var}(u' \lambda|u))$, since $E(u' \lambda|u) = 0$. Here, var($u' \lambda|u$) = $u' \text{cov}(\lambda)u = \sigma^2_s u'u$. In the prior, $\lambda_1, \ldots, \lambda_m$ are i.i.d. with mean zero and variance $\sigma^2_s$. Thus, var($u' \lambda$) = $\sigma^2_s E(u'u)$. Under the set-up of our hierarchical model, $\sigma^2_s = \lambda^2 | p, \sigma^2_1$ $\sigma^2_e = \sigma^2_s + (1 - p)\sigma^2_0$. Conditioned on $p, \sigma^2_1, \sigma^2_e$, the expression of heritability is

$$h^2 = \frac{(\frac{1}{n} \sum_{i=1}^{n} u'u_i + \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i)}{\sigma^2_s}.$$ 

Next, we discuss an approach to estimating $E(u'u)$. It can be shown that $u_i = \frac{1}{\sqrt{n}} \hat{g}_i, i = 1, \ldots, n$ for the GWAS individuals are uncorrelated. We estimate $E(u'u)$ by $\frac{1}{n} \sum_{i=1}^{n} u_i' u_i = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i$, $\hat{g}_i = W'x_i$, where $x_i$ denotes the genotype vector for all the local SNPs for ith individual. Note that, $\frac{1}{n} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i = \frac{1}{n} \sum_{i=1}^{n} x_i' W' W x_i = \frac{1}{n} \sum_{i=1}^{n} \text{trace}(x_i' W' W x_i) = \frac{1}{n} \sum_{i=1}^{n} \text{trace}(W' W x_i' x_i) = \frac{1}{n} \sum_{i=1}^{n} \text{trace}(W' W V) = \frac{1}{n} \sum_{i=1}^{n} \text{trace}(W' W V) = \frac{1}{n} \sum_{i=1}^{n} \text{trace}(W' V W)$. $V$ is the LD matrix in GWAS data. We denote $A = \frac{1}{n} \text{trace}(W' V W)$. Based on each MCMC sample of $p, \sigma^2_1, \sigma^2_e$, we also obtain a posterior sample of $h^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i$. We consider the posterior mean of $h^2$ as the estimated heritability and the posterior standard deviation as a measure of uncertainty.
2.9.1 | Choice of the hyperparameters

We adopt a MOM approach to choosing the hyperparameters in the prior distributions of \( p \) and \( \sigma^2 \). To estimate gene-level polygenicity, we focus on the TWAS data obtained from the tissue type that is most relevant for the trait. However, if expression data are available for other tissues in the reference panel (e.g., GTEx data), we can obtain TWAS statistics for different tissue types. We consider the trait’s TWAS data from a closely relevant tissue to estimate the hyperparameters. For example, if we consider the brain the primary tissue type of interest for BMI, we can use the adipose-specific TWAS statistics to estimate the hyperparameters. This is a partially empirical Bayes approach because we use the same GWAS data while computing the TWAS statistics for both tissue types.

The two shape parameters in the Beta prior of \( p \) are \( a_1, a_2 \), and the shape and scale parameters in the IG prior of \( \sigma^2 \) are \( b_1, b_2 \). For \( j \)th gene, we integrate out \( \lambda_j \) to obtain the distribution of the TWAS statistics conditioned only on \( p, \sigma^2 \):

\[
\mathcal{Z}_j \sim pN(0.1 + \sigma^2) + (1 - p)N(0, 1).
\]

Note that the variance of \( \mathcal{Z}_j \) corresponding to the spike component is \((1 \times 10^{-10})\), which we approximate as 1. We can derive the \( k \)th order raw population moment as \( E(\mathcal{Z}_j^k) = E_p E(\mathcal{Z}_j^k | p) \), where \( E(\mathcal{Z}_j^k | p) = E_{\sigma^2} E(\mathcal{Z}_j^k | p, \sigma^2) \). Here, \( E_{\sigma^2} E(\mathcal{Z}_j^k | p, \sigma^2) = p E(\mathcal{Z}_j^k | \mathcal{Z}_j \sim N(0, 1 + \sigma^2)) + (1 - p) E(\mathcal{Z}_j | \mathcal{Z}_j \sim N(0, 1)). \)

The odd order moments of \( \mathcal{Z}_j \) are zero because the distribution of \( \mathcal{Z}_j \) is symmetric at zero conditioned on \( p, \sigma^2 \). Thus, \( E(\mathcal{Z}_j) = E(\mathcal{Z}_j^3) = 0 \). We obtain the second and fourth-order moments of \( \mathcal{Z}_j \) as the following:

\[
E(\mathcal{Z}_j^2) = 1 + \frac{a_1}{a_1 + a_2} \frac{b_2}{b_1 - 1},
\]

\[
E(\mathcal{Z}_j^4) = 3 \left[ \frac{a_2}{a_1 + a_2} + \frac{a_1}{a_1 + a_2} \left( 1 + \frac{2b_2}{b_1 - 1} + \frac{b_2^2}{(b_1 - 1)(b_1 - 2)} \right) \right].
\]

Using the MOM approach we equate \( E(\mathcal{Z}_j^2) = m_2 \) and \( E(\mathcal{Z}_j^4) = m_4 \), where \( m_2 \) and \( m_4 \) are the second and fourth order sample raw moments: \( m_2 = \frac{1}{M} \sum_{j=1}^{M} \mathcal{Z}_j^2 \) and \( m_4 = \frac{1}{M} \sum_{j=1}^{M} \mathcal{Z}_j^4 \), where \( M \) is the total number of genes on all chromosomes.

Since we have two equations in four unknowns, we fix the values of \( a_1 \) and \( b_1 \) as \( a_1 = 0.1, b_1 = 3 \). Then we solve the equations to obtain the choices of \( a_2 \) and \( b_2 \). It is challenging and tedious to obtain the sixth and eighth-order moments and finally solve for all four unknowns.

Next, we discuss a strategy for choosing the hyperparameters \( d_1, d_2 \) in the prior of \( \sigma^2 \): \( \sigma^2 \sim IG(d_1, d_2) \). \( E(\sigma^2) = \frac{d_1}{d_1 - 1} \) and s.d.(\( \sigma^2 \)) = \( \frac{d_2}{(d_1 - 1)(d_1 - 2)} \). We assume that var(\( \mathcal{Y} \)) = 1 and the heritability due to predicted expressions of the genes, \( h^2 \), varies in the range 1%-50%, hence \( \sigma^2 \in (0.5 - 0.99) \). We choose \( d_1 = 35, d_2 = 25 \) which induces \( \sigma^2 \) belong to \( (0.5 - 0.99) \) with 95% probability and the prior mean of \( \sigma^2 \) to be the middle of the interval.

3 | RESULTS

3.1 | Simulation study

We perform an extensive simulation study to evaluate the efficiency of \textit{genepoly} concerning estimating the gene-level polygenicity and the true subset of non-null genes. We use the actual genotype data of 337K white-British individuals in the UK Biobank (UKBB) for our simulations.

3.1.1 | Simulation design

The software package Fusion\(^7\) analyzed the expression-genotype data in the Young Finish Sequencing (YFS) study to identify all the locally heritable genes in the whole blood tissue. Out of 4700 locally heritable genes, we considered a subset of 2988 in chromosomes 7-22 for our simulations. For each gene, we consider the same set of local SNPs that Fusion included to estimate the local heritability and the prediction model of the genetic component of expression. We consider a subset of \( n_F \) (eg, 1000) individuals randomly drawn from the UKBB individuals as the reference panel of expression data.

We use a linear model for each gene to simulate the expression in the reference panel: \( E_j = x_j' \mathbf{w}_j + e_j \). Here \( E_j \) denotes the expression of the \( j \)th gene, \( x_j \) denotes the genotypes of the set of local SNPs, and \( \mathbf{w}_j \) denotes the corresponding effect sizes.
Under the assumption that $V(E_j) = 1$, we consider $e_j \sim N(0, 1 - h^2_{ij})$, where $h^2_{ij}$ is the local heritability of the expression of $j$th gene (due to the local SNPs). We assume that a proportion of the local SNPs affect the expression. If $r_{ij}$ denotes the number of such SNPs, each element of $w_j$ follows $N(0, \frac{h^2_{ij}}{r_{ij}})$. While simulating the expression, we standardize the genotype data of each local SNP to have zero mean and unity variance. We apply Fusion to simulated data to identify the genes with significant local heritability using a $P$-value threshold $0.05$.

To evaluate this limitation’s impact on the number of non-null genes in GWAS, we ignore the uncertainty of the predicted genetic component of expression. We consider a subset of LD blocks randomly selected from a given chromosome and the genes with significant local heritability using a $P$-value threshold for the expression and local SNPs’ genotypes. We only consider the locallyheritable genes.

We next simulate the GWAS trait. We randomly select $n$ (e.g., 30,000) individuals from UKBB to create the GWAS cohort. We consider the reference panel and GWAS cohort individuals to be non-overlapping. If $p$ is the proportion of non-null genes among $m$ genes, we consider a subset of $m_c = [mp]$ genes to have a nonzero effect on the trait. Suppose expression data is not available in the GWAS. Assuming that the effect size of local SNPs on expression remains the same between the reference panel and the GWAS populations, we denote the true genetic component of the expression of $j$th non-null gene in GWAS as $G_j = X'_jw_j$, where $X_j$ denotes the genotype vector of the local SNPs in GWAS data, and $w_j$ denotes the true effect of local SNPs on the expression of $j$th non-null gene, $j = 1, \ldots, m_c$. We simulate $y = \sum_{j=1}^{m_c} G_j \alpha_j + e = G_c \alpha + e$, where $G_c = (G_1, \ldots, G_{m_c})'$ denotes the vector of the true genetic component of expressions for the non-null genes for an individual. Denote the matrix of true effect sizes of local SNPs on the expression of the non-null genes as $W_c$. In the GWAS data, if $X_c$ denotes the genotype matrix of the local SNPs for the non-null genes, $G_c = X_c W_c$ denotes the true genetic component of expression for the non-null genes. Thus, for $n$ individuals in the GWAS, we consider the simulation model as $Y = G_c \alpha + e$. Denote $M_{2c} = \sqrt{(\text{diag}(W'_c V_c W_c))^{-1}}$, where $V_c$ is the LD matrix of the local SNPs for the non-null genes in GWAS data. To normalize $G_c$ while simulating $Y$, we consider the transformation $G_c M_{2c}$. So, $Y = G_c M_{2c} \alpha + e = G_c M_{2c} \gamma + e$, where $\gamma = M_{2c}^{-1} \alpha$. Assume that $V(Y) = 1$ and the total heritability of $Y$ due to the genetic components of expressions of $m_c$ non-null genes is $h^2_{ij}$. Then the random noise $e \sim N(0, 1 - h^2_j)$. We simulate the elements of $\gamma = (\gamma_1, \ldots, \gamma_{m_c})$ independently: $\gamma_j \sim N(0, \frac{h^2_j}{m_c})$, $j = 1, \ldots, m_c$. After simulating $\gamma$, we obtain $\alpha = M_{2c} Y$.

We choose the number of non-null genes per chromosome proportional to the number of locallyheritable genes. On each chromosome, we consider the approximate LD blocks identified by Berisa and Pickrell such that each block contains at least one heritable gene. We consider a subset of LD blocks randomly selected from a given chromosome and randomly choose a gene from each block to be non-null. In the standard TWAS, we ignore the uncertainty of the predicted genetic component of expression. To evaluate this limitation’s impact on genepoly’s performance, we also performed a TWAS using the actual genetic component of expression in the second-stage regression based on the GWAS data. We plug in the actual effect sizes of the local SNPs on expression while computing the predicted genetic component of expression in the GWAS data. We refer to this approach as the benchmark TWAS. We compare the performance of genepoly applied to the standard and benchmark TWAS statistics. For a given dataset, we estimate the hyperparameters to be used in genepoly based on another dataset generated under the same simulation scenario using the MOM approach discussed above.

In the reference panel, we choose the local heritability of a gene at random between 10% and 15%. We consider 10% of the local SNPs to have a nonzero effect on the expression. We also consider a maximum of 300 SNPs for each gene. While simulating the GWAS trait, we consider a proportion of locallyheritable genes to have a non-null effect on the trait. We chose five different values of $p = 2\%, 4\%, 6\%, 8\%, 10\%$. Thus, for 2988 genes considered on chromosomes 7-22, the maximum number of non-null genes is considered to be 300. We consider two different choices of the sample size of the reference panel $n_E = 1000, 4000$, and the sample size of the GWAS data as $n = 30,000$ (30K), 50,000 (50K) (four different combinations of $n_E$ and $n$). We consider two different scenarios of trait heritability due to predicted expressions. In the first scenario, the heritability increases with the proportion of non-null genes, 10%, 12%, 14%, 16%, 18% heritability due to 2%, 4%, 6%, 8%, 10% non-null genes. In the second scenario, we fix the heritability at 10% and 20%. We run the benchmark TWAS for $n_E = 1000$ and $n_{GW} = 30K$.

3.1.2 Simulation results

Recall that $n_E$ and $n_{GW}$ denote the sample size of the reference panel expression and GWAS data, respectively; $p$ denotes the true proportion of non-null genes; $h^2$ denotes the trait heritability due to the genetic component of expressions. We measure the bias in estimation of $p$ using the relative bias $\frac{\text{estimated } p - \text{true } p}{\text{true } p}$. In simulation scenarios when $h^2$ increases proportionally with $p$ for $n_{GW} = 30K$ and $n_E = 1000, 4000$, genepoly provides an overall accurate estimates
of \( p \) (Figure 1a, Table 1). For \( p = 10\% \), \( h^2 = 18\% \), and \( n_E = 1000 \), \textit{genepoly} produces a downward bias which marginally improves by increasing \( n_E \) from 1000 to 4000.

Next, we discuss the results when \( p \) increases for a fixed \( h^2 \). When \( h^2 = 10\% \) and \( p = 2\% \), \textit{genepoly} produces an upward bias (Table 1), which may be explained by a larger value of mean \( h^2 \) per non-null gene. For a middle range of values of \( p \) (4\%, 6\%) considered, the method produced a small bias (Figure 1B, Table 1). As \( p \) increases to 8\%, 10\%, the mean heritability per non-null gene becomes smaller, and \textit{genepoly} underestimates \( p \) (Figure 1B, Table 1). The downward bias reduces marginally with an increase of \( n_E \) (1000-4000) in these simulation scenarios. For example, when \( h^2 = 10\% \) and \( p = 10\% \) with \( n_{GW} = 30K \), the mean relative bias is \(-36\% \) for \( n_E = 1000 \) compared to \(-31\% \) for \( n_E = 4000 \) (Table 1). A possible explanation is that the noise in the predicted expression decreases with an increase of \( n_E \). Encouragingly, the downward

FIGURE 1 Box plots for estimated polygenicity when trait heritability due to the genetic component of expressions are (A) 10\%, 12\%, 14\%, 16\%, 18\% for 2\%, 4\%, 6\%, 8\%, 10\% non-null genes, respectively, (B) fixed at 10\%, (C) fixed at 20\%. We consider two different choices of the sample size of the expression panel data \( (n_E) \) as 1000 and 4000. Here, the GWAS sample size is 30K.
TABLE 1 Percentage of relative bias produced by *genepoly* while estimating $p$ using standard TWAS and benchmark TWAS in various simulation scenarios.

| $h^2$  | $p$  | Standard TWAS | Benchmark TWAS |
|-------|------|---------------|----------------|
|       |      | $(n_E, n_{GW})$ | $(n_E, n_{GW})$ |   |
|       |      | 1000, 30K     | 1000, 50K      | 4000, 30K | 4000, 50K | 1000, 30K |
| 10%   | 2%   | 40 (23)       | 76 (26)        | 41 (27)    | 81 (30)  | −6 (15)  |
| 10%   | 4%   | 8 (14)        | 38 (17)        | 7 (14)     | 34 (13)  | −25 (9)  |
| 10%   | 6%   | −13 (10)      | 12 (13)        | −13 (11)   | 15 (12)  | −38 (7)  |
| 10%   | 8%   | −27 (9)       | −3 (11)        | −24 (9)    | 1 (10)   | −47 (5)  |
| 10%   | 10%  | −36 (7)       | −13 (9)        | −31 (7)    | −8 (9)   | −54 (5)  |
| 12%   | 4%   | 15 (14)       | 45 (16)        | 17 (14)    | 45 (14)  | −17 (10) |
| 14%   | 6%   | 3 (13)        | 34 (15)        | 4 (11)     | 33 (14)  | −26 (8)  |
| 16%   | 8%   | −3 (10)       | 24 (9)         | −2 (9)     | 25 (11)  | −28 (7)  |
| 18%   | 10%  | −8 (8)        | 17 (7)         | −5 (8)     | 23 (11)  | −31 (6)  |
| 20%   | 2%   | 90 (30)       | 129 (30)       | 87 (31)    | 133 (37) | 30 (20)  |
| 20%   | 4%   | 47 (17)       | 83 (20)        | 47 (17)    | 79 (16)  | 5 (12)   |
| 20%   | 6%   | 24 (10)       | 52 (16)        | 23 (12)    | 55 (15)  | −10 (8)  |
| 20%   | 8%   | 6 (9)         | 36 (14)        | 9 (10)     | 39 (12)  | −20 (7)  |
| 20%   | 10%  | −4 (9)        | 23 (10)        | 0 (8)      | 29 (12)  | −27 (6)  |

Note: $n_E$ and $n_{GW}$ denote the sample sizes of the reference panel expression and the GWAS data, respectively. The mean and s.d. of the estimates in a simulation scenario are provided.

bias diminishes rapidly with an increase of $n_{GW}$ (Figure 2A). For example, when $h^2 = 10\%$ and $p = 8\%$ with $n_E = 1000$, the mean relative bias reduced from $-27\%$ to $-3\%$ as $n_{GW}$ increased from 30K to 50K. When $h^2 = 20\%$ with other parameters are unchanged, the heritability per non-null gene increases. In these simulation settings, *genepoly* produces upward bias (Figure 2B,D). As expected, the upward bias decreases as the true value of $p$ increases (Table 1). Increasing $n_E$ did not help to reduce this bias in these settings (Table 1). We also note that in realistic scenarios, the heritability of a trait due to the genetic component of expressions is limited and not expected to be greater than 20\%. The q-value approach, which only takes $P$-values as the input and ignores any possible covariance structure of TWAS statistics, produced a huge upward bias while estimating $p$ for most of the simulation scenarios (Table S5).

In standard TWAS, we ignore the uncertainty of predicted expression. To explore the effect of this limitation, we also ran *genepoly* for benchmark TWAS, plugging in the true effects of local SNPs on the expression in the second-stage regression. We observe that *genepoly* based on the benchmark TWAS consistently underestimates $p$, both in the scenarios of $h^2$ increasing with $p$ and remaining fixed regardless of $p$ (Figure 3A-C, Table 1). It produces a slight upward bias in a few cases, for example, when $h^2 = 20\%$ and $p = 2\%, 4\%$. The overall results indicate that the benchmark TWAS framework is underpowered in general.

Besides investigating the performance of *genepoly* concerning the estimation bias, we also estimate the posterior standard deviation as the square root of the mean posterior variance obtained across the iterations in a given simulation scenario. We find that the posterior s.d. marginally increases as the true value of $p$ increases for a fixed choice of $h^2$ (Table S2). We also explore how the square root of the mean squared error (MSE) of estimated $p$ changes with the sample size variations (Table S3). As the expression panel data sample size increases, the square root of MSE decreases marginally in most cases. In some cases, it remains unchanged. As the GWAS sample size increases, it increases marginally in most simulation settings, while in some cases, it decreases (Table S3). Thus, we do not observe a specific pattern of the MSE for the GWAS sample size changes. *genepoly* estimates the total heritability due to predicted expressions with reasonably good accuracy (Figure 4, Table S1). Overall, the method estimates $p$ reasonably well, and we can improve the estimation accuracy by increasing the sample sizes of the reference panel expression and GWAS data in realistic simulation scenarios.

Next, we discuss the usefulness of *genepoly* while identifying the true subset of non-null genes. We measure the selection accuracy by specificity and sensitivity, where specificity measures the proportion of null genes excluded from the
FIGURE 2 Estimated polygenicity when the GWAS sample size increases from 30K to 50K for a fixed choice of the reference panel expression data sample size (n_E). In (A) and (B), we fix n_E = 1000 and consider the trait heritability due to the genetic component of expressions as 10% (A) and 20% (B), respectively. In both scenarios, we increase the GWAS sample size from 30K to 50K. We set n_E = 4000 and repeat the same analyses in (C) and (D).

Inferred subset, and sensitivity measures the proportion of non-null genes included in the subset. *genepoly* produces a very high level of specificity (mean specificity ≥ 93%) consistently across the various simulation scenarios (Table 2). It produces a decent overall sensitivity across the scenarios (Table 3). The mean sensitivity is higher when the mean h^2 per non-null gene is higher. For n_E = 1000 and n_GW = 30K, when h^2 = 20% with p = 2%, the mean sensitivity is 74% compared to 63% when h^2 = 10% (Table 3). Thus, the sensitivity decreases with the decrease in the mean heritability per non-null gene.

Sensitivity increases as n_E and n_GW increase. For n_E = 1000, the mean sensitivity increases when n_GW increases from 30K to 50K. For example, for h^2 = 10% and p = 4%, the mean sensitivity increases from 54% to 63% as n_GW increases from 30K to 50K (Table 3). Also, for a given choice of n_GW, the mean sensitivity increases when n_E increases. For example, for n_GW = 30K, h^2 = 10% and p = 6%, mean sensitivity increases from 46% to 52% when n_E increases from 1000 to 4000 (Table 3). A joint increase in n_E and n_GW provides the maximum sensitivity. For example, for h^2 = 10% and p = 2%,
FIGURE 3  Comparison between estimated polygenicity obtained by genepoly using the benchmark TWAS and standard TWAS when the trait heritability due to genetic component of expressions are (A) 10%, 12%, 14%, 16%, 18% for 2%, 4%, 6%, 8%, 10% non-null genes, respectively, (B) fixed at 10%, (C) fixed at 20%. The expression panel and GWAS data sample size are 1000 and 30K, respectively.

genepoly produces a maximum of 76% mean sensitivity (Table 3) when both the choices of $n_E$ and $n_{GW}$ are largest ($n_E = 4000$ and $n_{GW} = 50K$). genepoly based on the benchmark TWAS consistently produces a marginally higher (1%-3%) specificity compared to standard TWAS (Table 2) at the expense of a marginally lower sensitivity (1%) in a few simulation scenarios (Table 3). We also provide the ROC curves for some simulation scenarios (Figure 5). As expected, the AUC is higher for a larger $n_{GW}$ value. Thus, genepoly accurately identifies the true subset of non-null genes.

3.2  Real data application

We applied genepoly to seven phenotypes in UK Biobank (UKBB), three anthropometric traits, three lipid traits, and a case-control trait. The gene expression prediction model, estimated based on individual-level reference panel expression data, is required to implement genepoly. In the Fusion software package, which was mainly developed to implement
FIGURE 4  Comparison between estimated total heritability due to genetic component of expressions obtained by genePoly using the benchmark TWAS and standard TWAS when the trait heritability due to genetic component of expressions are fixed at (A) 10% and (B) 20%, respectively, for 2%, 4%, 6%, 8%, 10% non-null genes. The expression panel and GWAS data sample size are 1000 and 30K, respectively.

TABLE 2  Mean specificity of genePoly while inferring the subset of non-null genes using standard and benchmark TWAS in various simulation scenarios.

| $h^2$ | $p$ | Standard TWAS | Benchmark TWAS |
|-------|-----|---------------|----------------|
|       |     | ($n_E, n_{GW}$) | ($n_E, n_{GW}$) |
|       |     | 1000, 30K | 1000, 50K | 4000, 30K | 4000, 50K | 1000, 30K |
| 10%   | 2%  | 98 (0) | 97 (1) | 98 (1) | 98 (1) | 99 (0) |
| 10%   | 4%  | 98 (0) | 97 (1) | 98 (1) | 97 (1) | 99 (0) |
| 10%   | 6%  | 97 (1) | 96 (1) | 98 (1) | 97 (1) | 99 (0) |
| 10%   | 8%  | 97 (1) | 96 (1) | 97 (1) | 96 (1) | 99 (0) |
| 10%   | 10% | 97 (1) | 96 (1) | 97 (1) | 96 (1) | 99 (0) |
| 12%   | 4%  | 97 (0) | 96 (1) | 97 (1) | 97 (1) | 99 (0) |
| 14%   | 6%  | 97 (1) | 95 (1) | 97 (1) | 96 (1) | 98 (0) |
| 16%   | 8%  | 96 (1) | 95 (1) | 96 (1) | 95 (1) | 98 (0) |
| 18%   | 10% | 95 (1) | 94 (1) | 96 (1) | 94 (1) | 98 (0) |
| 20%   | 2%  | 97 (1) | 96 (1) | 97 (1) | 96 (1) | 98 (0) |
| 20%   | 4%  | 96 (1) | 95 (1) | 97 (1) | 95 (1) | 99 (0) |
| 20%   | 6%  | 96 (1) | 94 (1) | 96 (1) | 95 (1) | 99 (0) |
| 20%   | 8%  | 96 (1) | 94 (1) | 96 (1) | 94 (1) | 98 (0) |
| 20%   | 10% | 95 (1) | 93 (1) | 95 (1) | 93 (1) | 98 (0) |

Note: $n_E$ and $n_{GW}$ denote the sample sizes of the reference panel expression and the GWAS data, respectively. The s.d. of the estimates in a simulation scenario is also provided.
TABLE 3  Mean sensitivity of genepoly while inferring the subset of non-null genes using standard and benchmark TWAS in various simulation scenarios.

| $h^2$ | $p$ | Standard TWAS | Benchmark TWAS |
|-------|-----|---------------|---------------|
|       |     | (n_E, n_GW)   | (n_E, n_GW)   |
|       |     | 1000, 30K     | 1000, 50K     | 4000, 30K     | 4000, 50K     | 1000, 30K     |
| 10%   | 2%  | 63 (6)        | 71 (5)        | 66 (6)        | 76 (7)        | 63 (6)        |
| 10%   | 4%  | 54 (5)        | 63 (5)        | 58 (5)        | 67 (5)        | 53 (5)        |
| 10%   | 6%  | 46 (4)        | 58 (4)        | 52 (4)        | 62 (4)        | 46 (4)        |
| 10%   | 8%  | 42 (4)        | 53 (4)        | 47 (4)        | 59 (4)        | 41 (3)        |
| 10%   | 10% | 38 (3)        | 50 (4)        | 44 (3)        | 55 (3)        | 37 (3)        |
| 12%   | 4%  | 58 (5)        | 66 (4)        | 62 (5)        | 70 (4)        | 58 (5)        |
| 14%   | 6%  | 54 (4)        | 64 (4)        | 58 (4)        | 68 (4)        | 53 (4)        |
| 16%   | 8%  | 53 (4)        | 63 (4)        | 57 (4)        | 67 (4)        | 52 (3)        |
| 18%   | 10% | 52 (3)        | 62 (3)        | 57 (3)        | 66 (3)        | 51 (3)        |
| 20%   | 2%  | 74 (6)        | 79 (4)        | 76 (6)        | 83 (5)        | 74 (6)        |
| 20%   | 4%  | 67 (5)        | 74 (5)        | 70 (4)        | 77 (4)        | 66 (5)        |
| 20%   | 6%  | 61 (4)        | 70 (4)        | 65 (4)        | 73 (4)        | 61 (4)        |
| 20%   | 8%  | 57 (3)        | 66 (3)        | 61 (4)        | 71 (3)        | 57 (3)        |
| 20%   | 10% | 53 (3)        | 63 (3)        | 59 (3)        | 68 (3)        | 53 (3)        |

Note: n_E and n_GW denote the sample sizes of the reference panel expression and GWAS data, respectively. The s.d. of the estimates in a simulation scenario is also provided.

FIGURE 5  Receiver's operating characteristic (ROC) curves presenting the accuracy of genepoly while classifying the non-null genes. We consider a single dataset simulated under the following three simulation scenarios: 2%, 4%, 6% polygenicity for 10% trait heritability due to the genetic component of expressions. While the sample size of the expression panel data is fixed at 1000, the sample sizes of GWAS data are considered 30K in the top row and 50K in the bottom row.
TABLE 4  Estimation of gene-level polygenicity by genepoly for seven phenotypes in UK Biobank integrating external expression panel data.

| Trait       | Estimated polygenicity | Tissue-type ($n_E$) |
|-------------|------------------------|--------------------|
|             | Posterior median       | 95% posterior interval | Primary                  | Secondary                  |
| Height      | 18.6%                  | 16.6%              | 20.4%                     | Muscle skeletal (361)       | Adipose subcutaneous        |
| BMI         | 6.9%                   | 5.5%               | 8.4%                      | Brain cerebellum (103)      | Adipose subcutaneous        |
| WHR         | 4.7%                   | 3.8%               | 5.6%                      | Adipose subcutaneous (298)  | Muscle skeletal              |
| HDL         | 4.8%                   | 3%                 | 6.9%                      | Liver (97)                  | Whole blood (338)            |
| LDL         | 2.6%                   | 1.4%               | 4.3%                      | Liver                      | Whole blood                  |
| Triglycerides | 3.5%             | 2.2%               | 5.6%                      | Liver                      | Whole blood                  |
| Asthma      | 0.5%                   | 0.3%               | 0.8%                      | Whole blood (YFS) (1264)    | Whole blood                  |

Note: We used the expression prediction models from the Fusion software package. The expression prediction model based on whole blood considered a primary tissue type for asthma, was collected from the YFS. All other expression prediction models used here were fitted in the GTEx study. The sample sizes of the expression data are also provided in the last two columns.

standard TWAS based on summary statistics, the estimated prediction models are available for different expression datasets. GTEx is one such dataset that contains gene expressions in various tissue types and genotype data of genome-wide SNPs. We use the estimated prediction models of gene expressions in different tissue types of interest in the GTEx data and other expression panel data provided by the Fusion software package. We analyzed chromosomes 1-22 together. For each trait, we considered a primary and a secondary tissue type. We use the TWAS statistics based on the second tissue type to implement the MOM approach to estimate the hyperparameters in the Bayesian framework of genepoly. We finally implement genepoly based on the trait’s TWAS data obtained from the primary tissue type using Elastic Net to estimate the prediction model for expression. Finucane et al. developed a novel method to identify the relevant tissue or cell types for a complex trait and reported the most appropriate tissue types for a collection of traits. Their findings guide the choice of the tissue types pertinent to the traits considered in our analysis. For example, while analyzing a lipid trait (LDL, HDL, or triglyceride), we considered the liver tissue primary and whole blood secondary tissue type. Similarly, for WHR, we considered adipose the primary tissue and muscle-skeletal as the second tissue type. We note that Finucane et al. reported these tissue types as significantly relevant for WHR. We considered the GTEx expression data for most traits, except whole blood in Young Finish Study (YFS) for asthma as the primary tissue type.

In our analysis, height appeared to be most polygenic (posterior median of $p$ as 19% with the 95% central posterior interval as (17%, 20%)) (Table 4). For the other two anthropometric traits, BMI is 7% polygenic with a posterior interval of 6%-8%, and WHR is 5% polygenic with a posterior interval of 4%-6% (Table 4). Among the lipids, HDL is the most polygenic (5%), and LDL is the least polygenic (3%) with triglycerides falling between (4%). The posterior intervals of $p$ overlap partially between HDL and LDL (Table 4). It suggests HDL is likely more polygenic than LDL and deserves further investigation. In our analysis, the case-control trait asthma was the least polygenic (0.5%).

For each trait analyzed, we also report the estimated subset of non-null genes identified by genepoly (Tables S6-S14). While the subset for height contains 245 genes (Tables S12-S14), the subset for LDL contains nine genes (Table S8). Many of the genes included in the subset for a trait were previously reported to be associated with the trait or other relevant traits. For each trait, we mention such a gene in the following. For asthma, QSOX1 on chromosome 1 (Table S6) has been reported to be associated with blood protein measurements. For lipid traits, CELSR2 on chromosome 1 (Table S7-S9) is previously known to be associated with lipids (numerous studies reported in EBI GWAS catalog). For BMI, PPP2R3A on chromosome 3 (Table S10) has been reported to be associated with BMI in the EBI GWAS catalog. For WHR, GRK4 on chromosome 4 (Table S11) is previously known to be associated with BMI-adjusted hip circumference. For height, PEX1 on chromosome 7 (Table S13) is known to be associated with height and other anthropometric traits.

To contrast the gene-level polygenicity with SNP-level polygenicity, we applied the method Genesis for the seven complex traits in the UK Biobank. Genesis preferred a three-component than a two-component mixture model based on model-selection criteria for all the traits except for BMI. Genesis estimates the number of susceptibility SNPs as a measure of SNP-level polygenicity. We also consider the ratio of the number of susceptibility SNPs to all SNPs (Table 5). Here, the anthropometric traits appeared to be the most polygenic as before, but with WHR leading the list, followed by height and BMI. However, the difference in the estimates is marginal (Table 5). In contrast, genepoly identified height as the
| Trait      | #Mixture components | #Susceptibility SNPs (SD) | Total #SNPs | SNP-level polygenicity (SD) |
|------------|----------------------|--------------------------|-------------|-----------------------------|
| Height     | 3                    | 17 827 (1390)            | 548 688     | 0.032 (0.003)               |
| BMI        | 2                    | 17 036 (1203)            | 550 805     | 0.031 (0.002)               |
| WHR        | 3                    | 22 285 (3012)            | 550 940     | 0.04 (0.005)                |
| HDL        | 3                    | 14 162 (1630)            | 550 220     | 0.026 (0.003)               |
| LDL        | 3                    | 7500 (1810)              | 550 641     | 0.014 (0.003)               |
| Triglycerides | 3                  | 13 333 (1507)            | 550 447     | 0.024 (0.003)               |
| Asthma     | 3                    | 2714 (1521)              | 551 036     | 0.005 (0.003)               |

most polygenic trait with a pronounced difference in the estimate. Similarly, as genepoly, asthma was identified to be the least polygenic trait by Genesis (Table 5). Again, HDL and triglycerides are found to be more polygenic than LDL. The Pearson correlation coefficient between seven traits’ SNP- and gene-level polygenicity was estimated to be 0.53. Overall, we observe similar patterns between the SNP- and gene-level polygenicity of the traits and a few patterns unique to each type of polygenicity.

4 | DISCUSSION

We propose a Bayesian approach genepoly to estimate the gene-level polygenicity for a complex trait under the framework of TWAS. genepoly simultaneously estimates the trait’s true subset of non-null genes. It explicitly accounts for the covariance structure between the genes. The method uses summary-level TWAS association statistics, which are often publicly available. Along with the point estimate of gene-level polygenicity, genepoly provides a posterior interval to assess the uncertainty.

The simulation study shows that genepoly performs well in estimating the proportion of non-null genes in realistic scenarios. The downward bias produced by genepoly when the heritability per non-null gene is small diminishes rapidly with the increase of the GWAS sample size (\(n_{GW}\)). We experimented with a maximum \(n_{GW}\) of 50K for computational convenience in simulations. With the sample size of contemporary GWASs available (eg, hundreds of thousands in UK Biobank), we anticipate that the downward bias will reduce substantially. In simulations, we considered the choices of \(p\) as 2%-10%. If we increase \(p\) for a fixed \(h^2\), the downward bias will increase due to a decrease in the mean trait heritability per non-null gene. For \(h^2\) proportionally increasing with \(p\), genepoly is expected to perform well. We note that we included genes on chromosomes 7-22 in simulations for computational feasibility. We expect the overall findings to be similar for chromosomes 1-22.

Standard TWAS can inflate the false positive rate in association testing, that is, inflated test statistics for some null genes, mainly due to the non-adjustment of variance of the predicted expression in the GWAS data. As a consequence, we observe more overestimation bias for a standard TWAS. An upward bias increases further as \(n_{GW}\) increases. The lack of uncertainty adjustment in the predicted expression can inflate the test statistics of some null genes more severely for a larger \(n_{GW}\), which leads to an increase in the upward bias. We note that we obtained the estimation bias in both positive and negative directions. In many simulation scenarios, the magnitude of downward bias reduces as \(n_{GW}\) increases. Generally, when the heritability due to per non-null gene is small in the more realistic scenarios, we observe a downward bias, which reduces as \(n_{GW}\) increases. In the benchmark TWAS, we plug in the true effects of local SNPs on the expression while computing the predicted expression in the GWAS data. As a result, test statistics for the null genes should be well-calibrated and not inflated. Benchmark TWAS should maintain the correct false positive rate in association testing. The downward bias of benchmark TWAS can be attributed to \(n_E\) and \(n_{GW}\), that is, it should decrease with increased sample sizes, which we observe in simulations.

Genepoly performs well while identifying the true subset of non-null genes. The estimated subsets are highly specific, with an overall good level of sensitivity. The sensitivity improves if the sample size of the reference panel expression and GWAS data increase. The specificity for benchmark TWAS is marginally higher than standard TWAS. Overall, the selection accuracy of genepoly is comparable between standard and benchmark TWAS. Thus, functional studies can further
investigate the subset of non-null genes estimated for a complex trait. Here, we examine more closely how useful the estimated subsets of non-null genes are. For \( p = 4\% \) and \( h^2 = 10\% \) with \( n_E = 1000 \) and \( n_{GW} = 30K \), \textit{genepoly} produces an average of 98% specificity and 54% sensitivity (Tables 2 and 3). Among 2988 genes considered with \( p = 4\% \), 120 genes have a nonzero effect, and the rest are null. The specificity and sensitivity values imply that the identified subset of genes includes 65 non-null genes and 57 null genes on average. This indicates that the estimated subset contains a good number of non-null genes, which is useful for follow-up studies. In this work, we mainly focused on complex traits for which numerous genes are expected to have small effects and hence a higher polygenicity. In future work, we plan to explore the performance of our approach for a Mendelian disorder that has large effects due to a very small number of genes.

SNP-level polygenicity is based on the framework of GWAS, and our notion of gene-level polygenicity is based on the framework of TWAS. GWAS and TWAS have similar goals of discovering novel genetic variants associated with complex traits but with substantial distinction. In GWAS, associated SNPs are assumed to have a marginal effect on the trait regardless of the underlying molecular mechanism. In TWAS, the impact of a gene’s local SNPs on the trait is assumed to be mediated through its expression. Not all GWAS susceptibility SNPs may affect the trait mediated through expression. Thus, SNP- and gene-level polygenicity are fundamentally different quantities.

While developing the Bayesian method, we also experimented with the Dirac spike and slab prior (point mass at zero). Our preliminary simulations showed that the MCMC implementation does not perform robustly, particularly for many genes (in thousands). We also found that the MCMC for the continuous spike and slab prior is computationally faster than for the Dirac spike and slab prior. We implemented a MOM approach to choose the hyperparameters in the model based on a tissue type closely relevant to the trait. We propose using the TWAS statistics from a tissue type nearly appropriate to the trait. If the data for a second tissue type is unavailable, we can always apply the MOM approach for the primary tissue (an entirely empirical Bayes approach). We have described \textit{genepoly} for a continuous trait. However, the method can also be applied to a binary trait, mainly because the form of the covariance matrix for marginal TWAS statistics depends on the expression prediction model obtained from the reference panel and the LD matrix of the eQTLs. \textit{Focus} is a fine-mapping method under the TWAS framework which also considers covariance between the TWAS statistics. We note that the analytical form of covariance is similar between \textit{focus} and \textit{genepoly}. This article uses a “non-null gene” instead of a “causal gene.” The genetic component of the expression of a non-null gene has a nonzero effect on the trait. Our definition does not imply a formally causal biological relationship between the non-null gene and the trait. In future work, we plan to estimate the gene-level polygenicity while adjusting for the uncertainty of predicted expression in TWAS.

\textit{genepoly} is a methodologically sound unified Bayesian approach to estimating the gene-level polygenicity and the subset of non-null genes for a complex trait under the TWAS framework. It is computationally efficient and can be applied to various traits to understand their genetic architecture better.

**ACKNOWLEDGEMENTS**

We sincerely thank Dr. Tanushree Halder for helping with the graphical presentation of the paper. We acknowledge Dr. Nicholas Mancuso for helpful discussions related to this work. This research used the UK Biobank resource under applications 33297 and 33127.

**DATA AVAILABILITY STATEMENT**

The datasets that we have analyzed or used in our analyses in this article are available (either openly or via applications) from the following websites: UK Biobank: https://www.ukbiobank.ac.uk/; GTEx: https://gtexportal.org/home/; Fusion: http://gusevlab.org/projects/fusion/.

**ORCID**

\textit{Arunabha Majumdar} \( \text{https://orcid.org/0000-0002-5609-075X} \)

**REFERENCES**

1. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. \textit{Nat Genet.} 2018;50(9):1318-1326.
2. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. \textit{Nat Genet.} 2013;45(4):400-405.
3. O’Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme polygenicity of complex traits is explained by negative selection. \textit{Am J Hum Genet.} 2019;105(3):456-476.
4. Johnson R, Burch KS, Hou K, Paciuc M, Pasaniuc B, Sankararaman S. Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits. *PLoS Comput Biol*. 2021;17(10):e1009483.

5. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019;51(4):592-599.

6. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098.

7. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245-252.

8. Li B, Ritchie MD. From GWAS to gene: transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. *Front Genet*. 2021;12:71320.

9. GTEx Consortium, Ardlie KG, Deluca DS, et al. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660.

10. Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.

11. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825.

12. Cao C, Ding B, Li Q, Kwok D, Wu J, Long Q. Power analysis of transcriptome-wide association study: implications for practical protocol choice. *PLoS Genet*. 2021;17(2):e1009405.

13. Siewert-Rocks KM, Kim SS, Yao DW, Shi H, Price AL. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *Am J Hum Genet*. 2022;109(3):393-404.

14. George EI, McCulloch RE. Variable selection via Gibbs sampling. *J Am Stat Assoc*. 1993;88(423):881-889.

15. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann Stat*. 2005;33(2):730-773.

16. Ročková V. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann Stat*. 2018;46(1):401-437.

17. Malsiner-Walli G, Wagner H. Comparing spike and slab priors for Bayesian variable selection. *Austrian J Stat*. 2016;40(4):241-264.

18. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol*. 1996;58(1):267-288.

19. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology*. 2005;67(2):301-320.

20. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213.

21. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat*. 2011;5(3):1780-1815.

22. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*. 2016;32(2):283.

23. Yao DW, O’Connor L, Price L, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet*. 2020;52(6):626-633.

24. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat*. 2003;31(6):2013-2035.

25. Finucane HK, Reshef YA, Anttila V, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet*. 2018;50(4):621-629.

26. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73-79.

27. Zhu Z, Guo Y, Shi H, et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK biobank. *J Allergy Clin Immunol*. 2020;145(2):537-549.

28. Justice AE, Young K, Gogarten SM, et al. Genome-wide association study of body fat distribution traits in Hispanics/Latinos from the HC/HS/SOL. *Hum Mol Genet*. 2021;30(22):2190-2204.

29. Tachmazidou I, Süveges D, Min JL, et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am J Hum Genet*. 2017;100(6):865-884.

30. Mancuso N, Freund MK, Johnson R, et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet*. 2019;51(4):675-682.

---

**How to cite this article:** Majumdar A, Pasaniuc B. A Bayesian method for estimating gene-level polygenicity under the framework of transcriptome-wide association study. *Statistics in Medicine*. 2023;42(26):4867-4885. doi: 10.1002/sim.9892