The American Journal of Human Genetics, Volume *109*

Supplemental information

A multi-dimensional integrative scoring framework

for predicting functional variants in the human genome

Xihao Li, Godwin Yung, Hufeng Zhou, Ryan Sun, Zilin Li, Kangcheng Hou, Martin Jinye Zhang, Yaowu Liu, Theodore Arapoglou, Chen Wang, Iuliana Ionita-Laza, and Xihong Lin

Supplemental Figures

Figure S1. Models representing potential causal relations among annotations. (a) All annotations $\boldsymbol{y}_{ij}$ (e.g., conservation measures, epigenetic measures) are treated as consequences of a single latent dichotomous variable of function $\boldsymbol{c}_i$. Annotations are assumed to be independent conditional on $\boldsymbol{c}_i$, as proposed in GenoCanyon. (b) All annotations $\boldsymbol{y}_{ij}$ are treated as consequences of $\boldsymbol{c}_i$. Annotations may be correlated conditional on $\boldsymbol{c}_i$. (c) There are multiple, possibly related, latent dichotomous variables of function $c_{i1}, \dots, c_{iM}$. For each functional status c_{ij}, a subset of annotations $y_{ij1}, \dots, y_{ijL_j}$ are observed as consequences. Annotations measuring the same c_{ij} may be correlated conditional on c_{ij}, as proposed in MACIE.

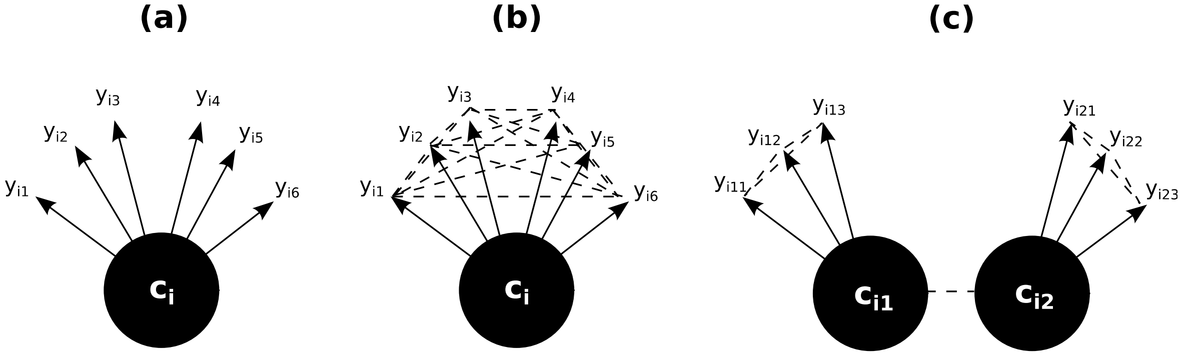


Figure S2. ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign missense variants.

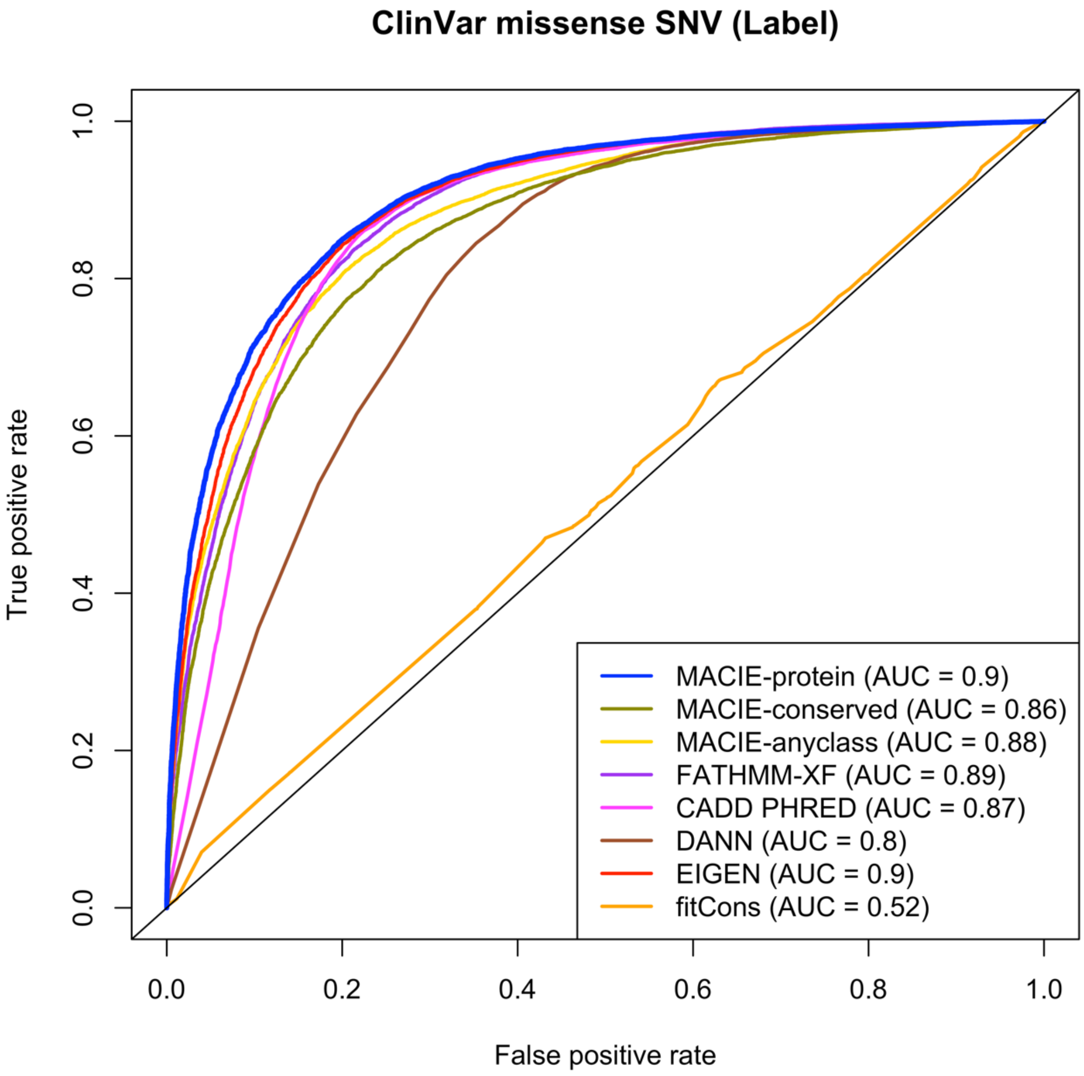


Figure S3. ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign non-coding variants.

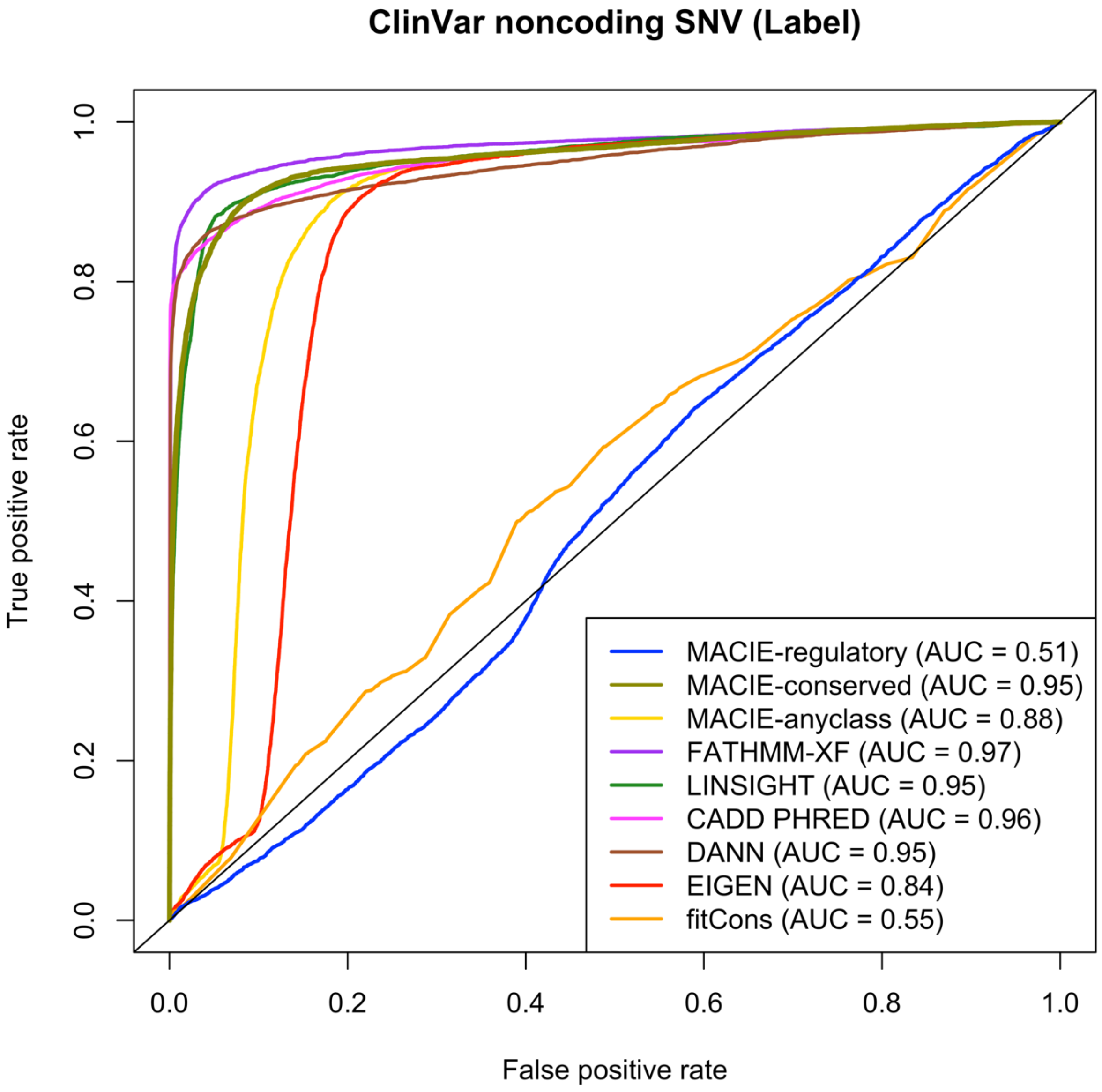

ClinVar noncoding SNV (Label)

Figure S4. ROC curves comparing the performances of MACIE and other functional scores in discriminating between loss-of-function (LOF) non-synonymous coding variants within 13 exons that encode functionally critical domains of *BRCA1* (putative functional class) based on saturation genome editing (SGE) data and ClinVar benign non-synonymous coding variants (putative non-functional class).

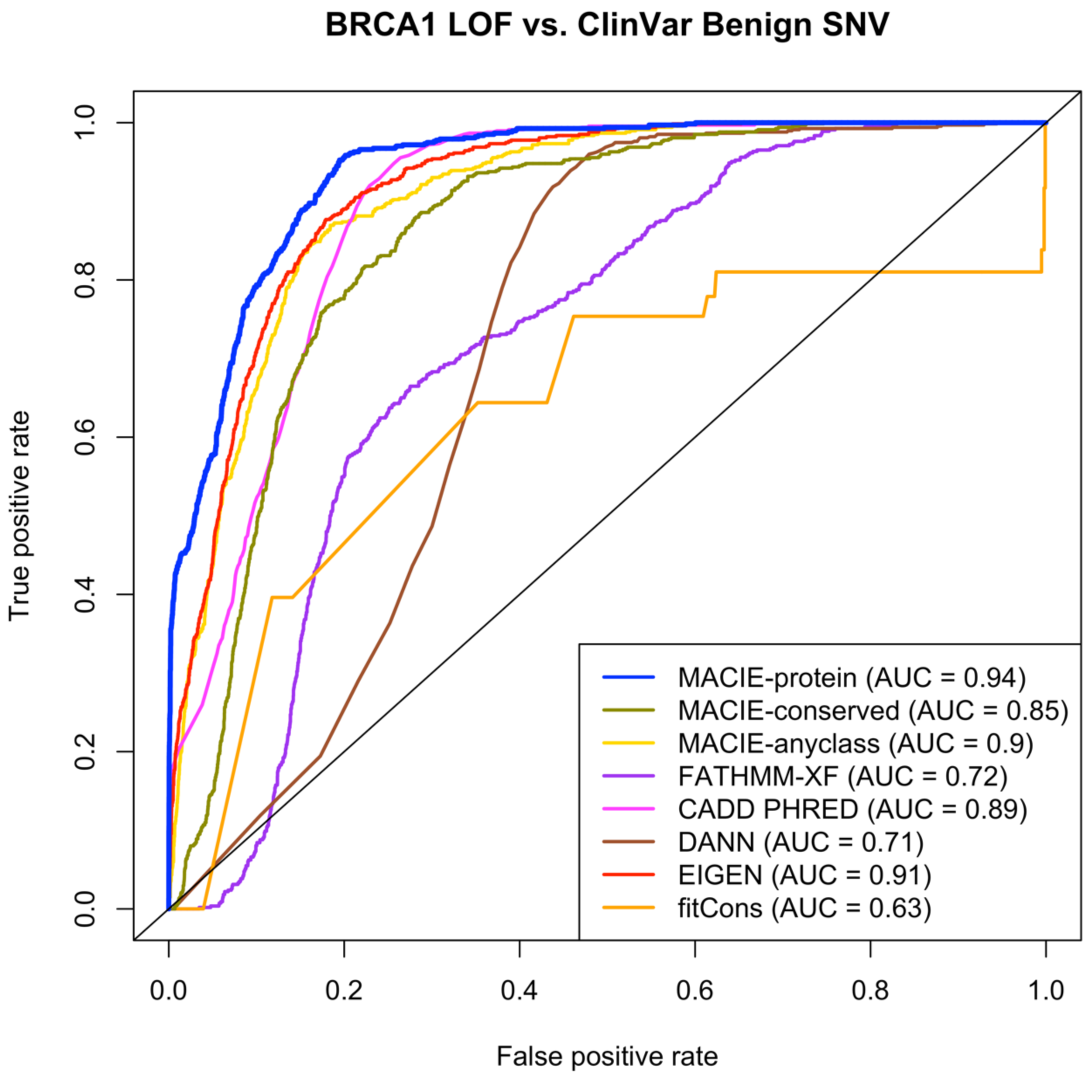


Figure S5. LocusZoom plot[1] for GWAS associations of TC at the *APOE* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (n = 62,166). The MACIE-protein and MACIE-conserved scores for rs7412 are 0.96 and 0.97, respectively. The MACIE-conserved and MACIE-regulatory scores for rs1065853 are < 0.01 and > 0.99, respectively. TC, total cholesterol.

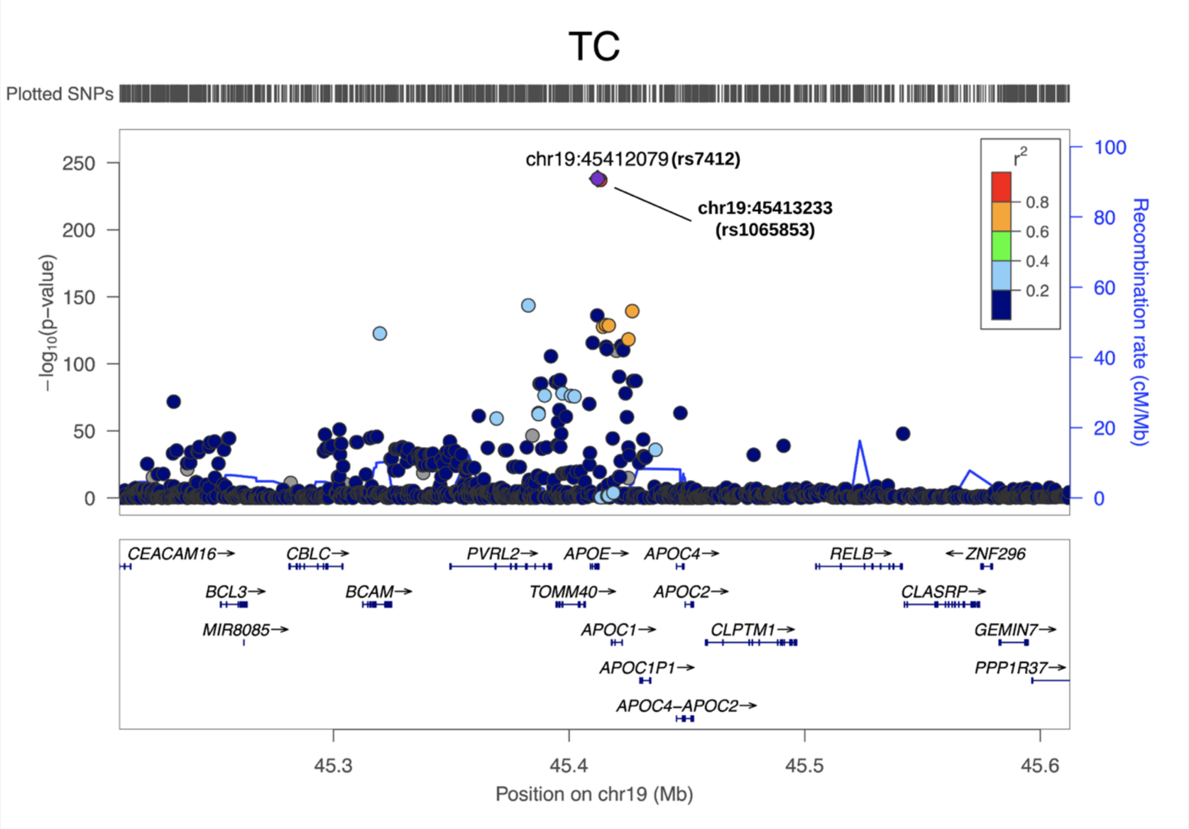


Figure S6. LocusZoom plot[1] for GWAS associations of HDL-C at the *CETP* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (n = 60,812). The MACIE-conserved and MACIE-regulatory scores for rs17231506 are both < 0.01. For both rs72786786 and rs12720926, the MACIE-conserved and MACIE-regulatory scores are < 0.01 and > 0.99, respectively. HDL-C, high-density lipoprotein cholesterol.

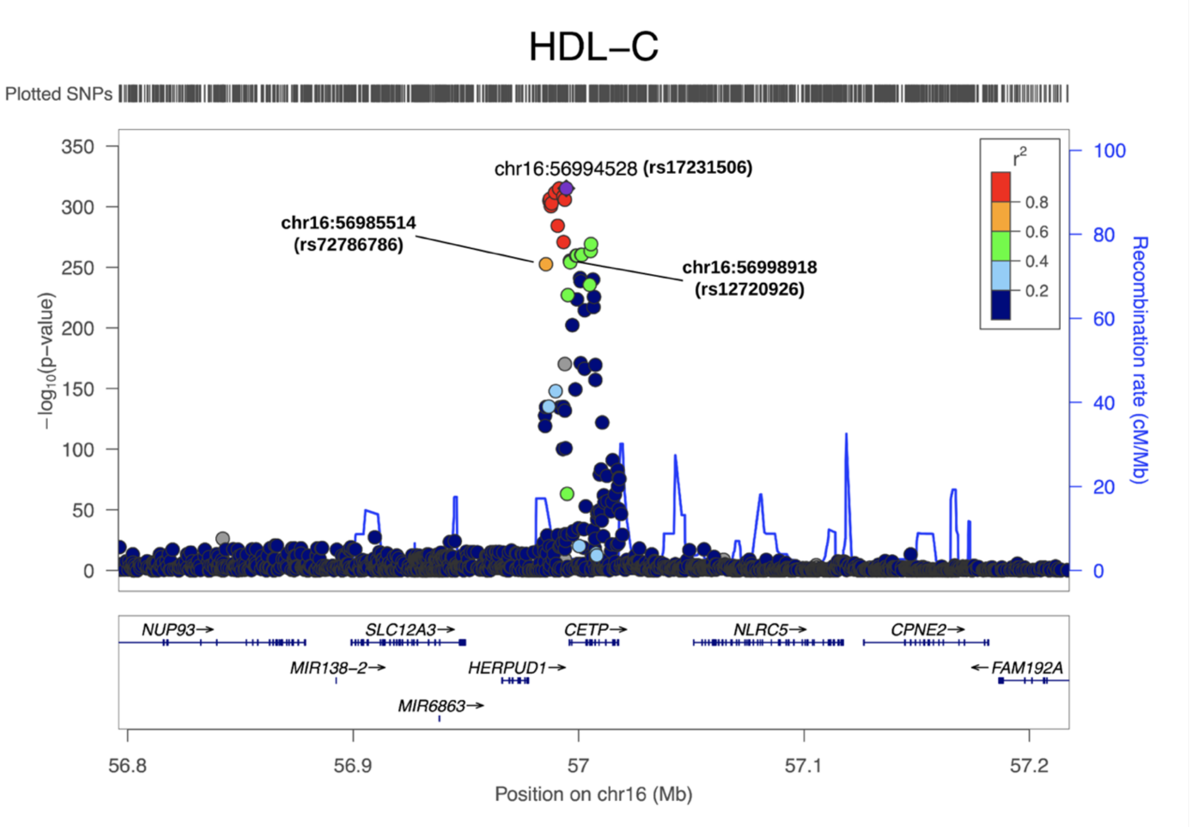


Figure S7. LocusZoom plot[1] for GWAS associations of TG at the *APOC3* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (n = 60,027). The MACIE-conserved and MACIE-regulatory scores for rs964184 are both < 0.01. The MACIE-conserved and MACIE-regulatory scores for rs2075290 are < 0.01 and 0.88, respectively. TG, triglycerides.

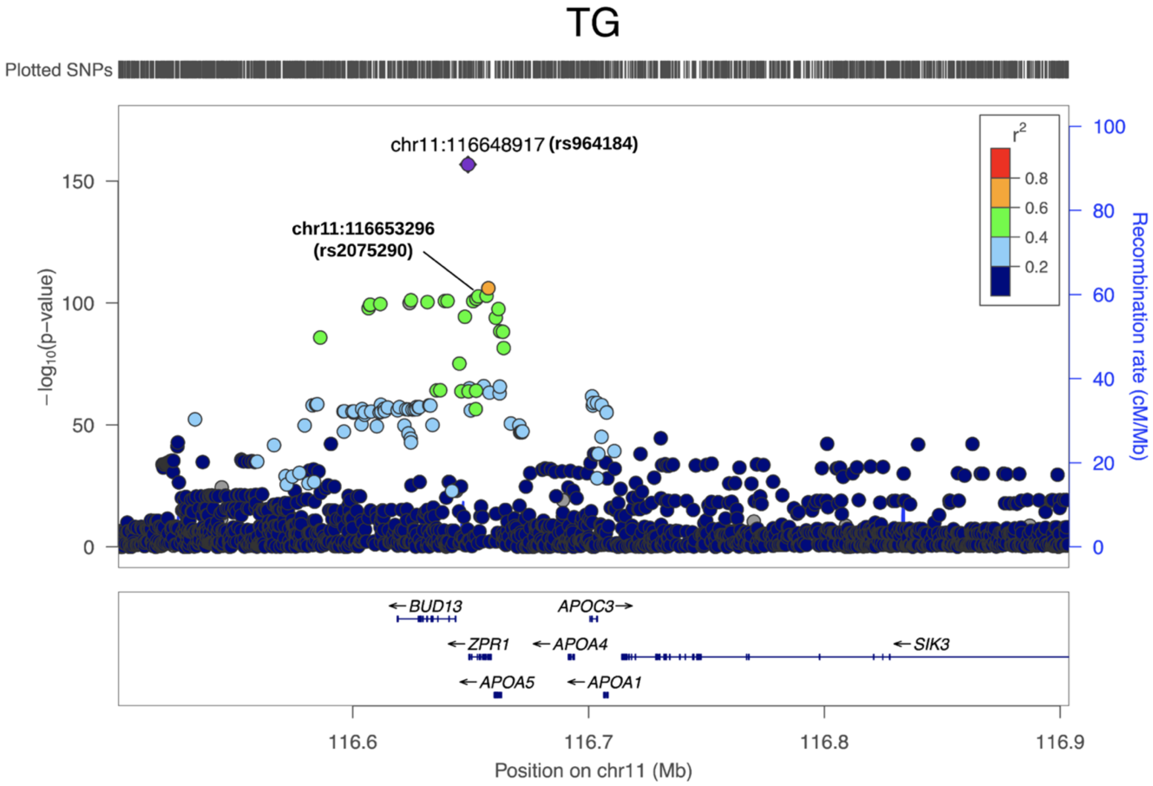
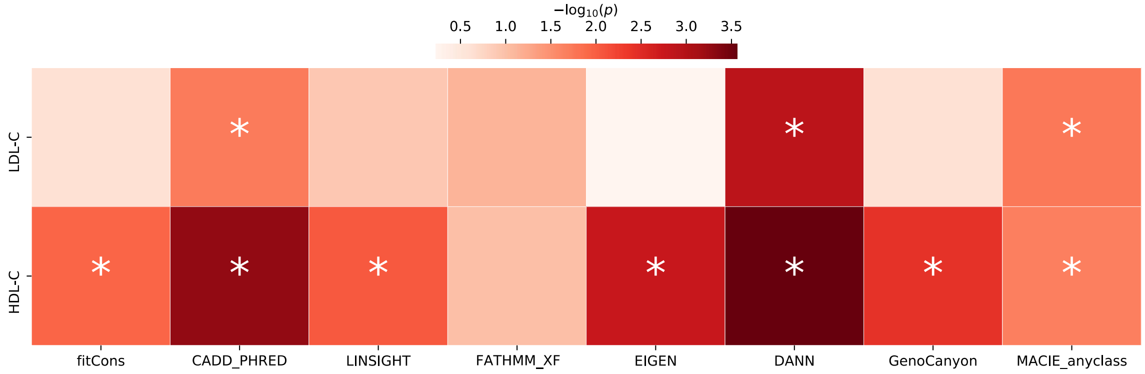


Figure S8. Heritability enrichment for 2 lipid traits from the ENGAGE Consortium, LDL-C and HDL-C, using 8 integrative annotations, fitCons, CADD, LINSIGHT, FATHMM-XF, EIGEN, DANN, GenoCanyon, and MACIE-anyclass. Asterisks indicate significant enrichment at FDR = 0.05 across all 16 trait-annotation pairs). LDL-C, low density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol.



Supplemental Material and Methods

The MACIE Generalized Linear Mixed Model (GLMM)

Suppose there are N genetic variants in total and we are interested in M latent annotation classes, each containing L_j annotation scores. For genetic variant i and annotation class j, we denote the set of L_j annotations as $\boldsymbol{y}_{ij} = \left(y_{ij1}, \dots, y_{ijL_j}\right)^T$, such that each variant is described by $L = \sum_{j=1}^{M} L_j$ annotations in total. We want to estimate for each variant i the vector of binary functional statuses $\boldsymbol{c}_i = (c_{i1}, \dots c_{iM})$, where c_{ij} is the unobserved latent functional status for class j. Conditional on c_{ij} and a random effect variable b_{ijk}, we assume that the elements of $\boldsymbol{y}_{ij}$ are independent observations, each generated from a one-parameter exponential family with canonical parameterization. That is, for $j = 1, \dots, M$ and $k = 1, \dots, L_j$,

$$f_{jk}\left(y_{ijk} | c_{ij}, b_{ijk}\right) = \exp\left[\frac{\{y_{ijk}\eta_{ijk} - d_{jk}(\eta_{ijk})\}}{\phi_{jk}} + h_{jk}\left(y_{ijk}, \phi_{jk}\right)\right], \quad (1)$$

with

$$\mu_{ijk} = E\left(y_{ijk}\right) = d'_{jk}\left(\eta_{ijk}\right),$$
$$V_{ijk} = \mathrm{Var}\left(y_{ijk}\right) = d''_{jk}\left(\eta_{ijk}\right)\phi_{jk},$$

where $\eta_{ijk} = g_{jk}\left(\mu_{ijk}\right)$ is a linear function of the functional status c_{ij} and random effect variable b_{ijk} such that

$$\eta_{ijk} = \beta_{0jk} + c_{ij}\beta_{1jk} + b_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_{jk} + b_{ijk}$$

for $\boldsymbol{x}_{ij} = \left(1, c_{ij}\right)^T$ and $\boldsymbol{\beta}_{jk} = \left(\beta_{0jk}, \beta_{1jk}\right)^T$. Additional correlations between elements of $\boldsymbol{y}_{ij}$ are allowed by assuming that

$$\boldsymbol{b}_{ij} = \begin{pmatrix} b_{ij1} \\ \vdots \\ b_{ijL_j} \end{pmatrix} \overset{iid}{\sim} MVN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_j(\boldsymbol{\theta})\right).$$

The marginal distribution of $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}^T, \dots, \boldsymbol{y}_{iM}^T)^T$ can be obtained by integrating over the distribution of $\boldsymbol{c}_i$ and $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \dots, \boldsymbol{b}_{iM}^T)^T$,

$$f(\boldsymbol{y}_i) = \sum_{c_{i1}=0,\dots,c_{iM}=0}^{1,\dots,1} \left(\prod_{j=1}^{M} \int f\left(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}\right) f\left(\boldsymbol{b}_{ij}, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{b}_{ij}\right) p(c_{i1}, \dots, c_{iM}). \quad (2)$$

Our primary focus concerns estimation of $p(\boldsymbol{c}_i | \boldsymbol{y}_i)$, the posterior probability of $\boldsymbol{c}_i$ conditional on the observed data $\boldsymbol{y}_i$. Because of the conditional independence of $\boldsymbol{y}_i$ given $\boldsymbol{c}_i$ and $\boldsymbol{b}_i$ (the collections of $\boldsymbol{y}_{ij}$, c_{ij}, and $\boldsymbol{b}_{ij}$ for $j = 1, \dots, M$, respectively), an

Expectation-Maximization (EM) algorithm provides a natural approach.[2] However, the integration in Equation (2) cannot be evaluated in closed form whenever $\boldsymbol{y}_{ij}$ conditional on c_{ij} and $\boldsymbol{b}_{ij}$ is not normally distributed (e.g. $\boldsymbol{y}_{ij}$ has dichotomous components). Thus, challenges arise in computing $p(\boldsymbol{c}_i|\boldsymbol{y}_i) = f(\boldsymbol{y}_i|\boldsymbol{c}_i)p(\boldsymbol{c}_i)/f(\boldsymbol{y}_i)$. Approximations are used when applying the EM algorithm to obtain parameter estimates for non-normally distributed annotations.

Given the fitted model parameters and the full set of annotation scores for a new genetic variant i', the MACIE score of variant i' is defined as the (predicted) posterior probability vector $\hat{p}(\boldsymbol{c}_{i'} = \boldsymbol{z}|\boldsymbol{y}_{i'}), \boldsymbol{z} \in \{0,1\}^M$. It can be calculated by performing one additional iteration of the EM algorithm.

Derivation of the EM Algorithm Used in MACIE GLMM

In the following, we let $\mathbf{1}_m$ be the vector of length m where each element takes the value 1, and let $\mathbf{J}_m$ be the $m \times m$ matrix of ones, i.e. $\mathbf{J}_m = \mathbf{1}_m \times \mathbf{1}_m^T$. Let $\mathbf{I}_m$ be the $m \times m$ identity matrix. Subscripts are dropped whenever the dimensions of the vector or matrix are clear. Our derivations follow those of Sammel et al.,[3] who considered a general class of latent variable models that allow for linear effects of covariates on multiple outcomes.

Maximization Step

If $\boldsymbol{c}_i$ and $\boldsymbol{b}_{ij}$ were directly observable, one can maximize the complete data log-likelihood,

$$\log f(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{b}) = \sum_{i=1}^{N} \left(\sum_{j=1}^{M} \sum_{k=1}^{L_j} \log f_{jk}\left(y_{ijk}|c_{ij}, b_{ijk}; \boldsymbol{\beta}_{jk}, \phi_{jk}\right) + \sum_{j=1}^{M} \log f\left(\boldsymbol{b}_{ij}; \boldsymbol{\theta}\right) + \log p(\boldsymbol{c}_i; \boldsymbol{\gamma}) \right) \tag{3}$$

to estimate the unknown parameters $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, where $\boldsymbol{\gamma}$ corresponds to the vector of length 2^M that holds the probability of each possible realization of $\boldsymbol{c}_i$. However, since $\boldsymbol{c}_i$ and $\boldsymbol{b}_{ij}$ are unobservable, the EM algorithm can be applied by instead solving the expected score functions, where the expectation is taken with respect to

$$f(\boldsymbol{c}_i, \boldsymbol{b}_i|\boldsymbol{y}_i) = f(\boldsymbol{b}_i|\boldsymbol{c}_i, \boldsymbol{y}_i)p(\boldsymbol{c}_i|\boldsymbol{y}_i) = \prod_{j=1}^{M} f\left(\boldsymbol{b}_{ij}|\boldsymbol{y}_{ij}, c_{ij}\right) \cdot p(\boldsymbol{c}_i|\boldsymbol{y}_i),$$

which is the posterior distribution of the missing data conditional on the observed data.[4] If we let $S_i(\boldsymbol{\zeta})$ denote the complete data score function $\partial \log f(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{b}_i) / \partial \boldsymbol{\zeta}$ of Equation

(3) for the ith variant, then each variant's contribution to the expected score function for $\boldsymbol{\gamma}$ is given by

$$E_{\boldsymbol{c},\boldsymbol{b}} S_i\left(\gamma_{z_1,\ldots,z_M}\right) = \frac{p(c_{i1}=z_1,\ldots,c_{iM}=z_M|\boldsymbol{y}_i)}{\gamma_{z_1,\ldots,z_M}} \tag{4}$$

for all $(z_1, \ldots z_M) \in \{0,1\}^M$. Therefore, based on Equation (4), the M-step update for $\boldsymbol{\gamma}$ in moving from iteration r to $r+1$ is

$$\hat{\gamma}^{(r+1)}_{z_1,\ldots,z_M} = \frac{\sum_{i=1}^{N} \hat{p}^{(r)}(c_{i1}=z_1,\ldots,c_{iM}=z_M|\boldsymbol{y}_i)}{N}. \tag{5}$$

For $\boldsymbol{\zeta}_{jk}$, the subset of parameters corresponding to only the jkth outcome, the contribution to the expected score equation for variant i is

$$E_{\boldsymbol{c},\boldsymbol{b}} S_i\left(\boldsymbol{\zeta}_{jk}\right) = \sum_{\boldsymbol{c}_i \in \{0,1\}^M}\left(\int S_i\left(\boldsymbol{\zeta}_{jk}\right) f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{c}_i)\mathrm{d}\boldsymbol{b}_i\right) p(\boldsymbol{c}_i|\boldsymbol{y}_i). \tag{6}$$

Depending on the form of the score function associated with the complete data log-likelihood $S\left(\boldsymbol{\zeta}_{jk}\right) = \sum_{i=1}^{N} S_i\left(\boldsymbol{\zeta}_{jk}\right)$, the solution to $E_{\boldsymbol{c},\boldsymbol{b}} S\left(\boldsymbol{\zeta}_{jk}\right) = \mathbf{0}$ may or may not be available in closed form. In the absence of a closed form solution, we update the estimates $\boldsymbol{\zeta}_{jk}$ through a one-step Fisher scoring algorithm. The usual method of estimation for this model is iteratively reweighed least squares,[5] where the weight function is updated at every iteration.

Expectation Step

Given the current estimates of the parameters, $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, the E-step is complicated by the need to compute expectations with respect to the posterior distributions $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{c}_i)$ and $p(\boldsymbol{c}_i|\boldsymbol{y}_i)$ of the missing data, conditional on the observed data. Only for normal outcomes will the posterior distributions have closed form solutions. In our setting, there are generally no closed form expressions for $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{c}_i)$ and $p(\boldsymbol{c}_i|\boldsymbol{y}_i)$. As an alternative to analytical solutions, we first write the expectation of functions of the data $g(\boldsymbol{c}_i, \boldsymbol{b}_i) = g(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{b}_i)$ by,

$$E_{\boldsymbol{c},\boldsymbol{b}}\, g(\boldsymbol{c}_i, \boldsymbol{b}_i) = \sum_{\boldsymbol{c}_i \in \{0,1\}^M}\left(\int g(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{c}_i)\mathrm{d}\boldsymbol{b}_i\right) p(\boldsymbol{c}_i|\boldsymbol{y}_i), \tag{7}$$

and further rewrite the posterior distributions as

$$f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{c}_i) = \prod_{j=1}^{M} \frac{f\left(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}\right) f\left(\boldsymbol{b}_{ij}\right)}{\int f\left(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}\right) f\left(\boldsymbol{b}_{ij}\right)\mathrm{d}\boldsymbol{b}_{ij}},$$

$$p(\boldsymbol{c}_i|\boldsymbol{y}_i) = \frac{\prod_{j=1}^{M}\left[\int f\left(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}\right) f\left(\boldsymbol{b}_{ij}\right)\mathrm{d}\boldsymbol{b}_{ij}\right] \cdot p(\boldsymbol{c}_i)}{\sum_{\boldsymbol{c} \in \{0,1\}^M} \prod_{j=1}^{M}\left[\int f\left(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}\right) f\left(\boldsymbol{b}_{ij}\right)\mathrm{d}\boldsymbol{b}_{ij}\right] \cdot p(\boldsymbol{c})}.$$

By substituting into Equation (7), we obtain

$$E_{\boldsymbol{c},\boldsymbol{b}} g(\boldsymbol{c}_i, \boldsymbol{b}_i) = \frac{\sum_{\boldsymbol{c}_i \in \{0,1\}^M} \left[\int g(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{b}_i) \prod_{j=1}^{M} f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_i \right] \cdot p(\boldsymbol{c}_i)}{\sum_{\boldsymbol{c}_i \in \{0,1\}^M} \prod_{j=1}^{M} \left[\int f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij} \right] \cdot p(\boldsymbol{c}_i)}. \tag{8}$$

If $g(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{b}_i) = g(\boldsymbol{y}_{ij'}, \boldsymbol{c}_{ij'}, \boldsymbol{b}_{ij'})$ for some $j' \in \{1, \dots, M\}$, then the integral in the numerator of Equation (8) is equivalent to

$$\prod_{j=1}^{M} \left[\int g(\boldsymbol{y}_{ij'}, \boldsymbol{c}_{ij'}, \boldsymbol{b}_{ij'})^{1(j=j')} f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij} \right] \tag{9}$$

where $1(j = j')$ is equal to 1 if $j = j'$ and 0 otherwise. In this case, a practical approach for approximation is to use multivariate Gauss-Hermite quadrature. To approximate Equation (9), we select T fixed abscissae $\{z_t\}_{t=1}^{T}$ and corresponding weights $\{w_t\}_{t=1}^{T}$ for a quadrature whose integration kernel is given by the density of a standard normal distribution.[6] Given the spectral decomposition of $\boldsymbol{\Sigma}_j = \mathbf{S}_j \boldsymbol{\Lambda}_j \mathbf{S}_j^T$, let $\sigma_{jt} = \{\sigma_{jt}(1), \dots, \sigma_{jt}(L_j)\}$ be an ordered set of L_j integers obtained by sampling with replacement from $\{1, \dots, T\}$, $\boldsymbol{z}_{jt} = \left(z_{\sigma_{jt}(1)}, \dots, z_{\sigma_{jt}(L_j)}\right)^T$ the corresponding set of abscissae, and $\boldsymbol{b}_{jt} = \mathbf{S}_j \boldsymbol{\Lambda}_j^{1/2} \boldsymbol{z}_{jt}$. Then each term in the product of Equation (9)

$$\int g(\boldsymbol{y}_{ij'}, \boldsymbol{c}_{ij'}, \boldsymbol{b}_{ij'})^{1(j=j')} f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij}$$

can be approximated as

$$\sum_{\sigma_{jt}} \left(\prod_{k=1}^{L_j} w_{\sigma_{jt}(k)} \right) g(\boldsymbol{y}_{ij'}, \boldsymbol{c}_{ij'}, \boldsymbol{b}_{j't})^{1(j=j')} f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{jt}),$$

where the sum is over all the possible ordered sets σ_{jt}. For some ordered sets σ_{jt} the weights $\prod_{k=1}^{L_j} w_{\sigma_{jt}(k)}$ are very small and thus contribute little to the sum. We may choose to remove these quantities by pruning a specified fraction of the smallest weights.

MACIE: EM Algorithm for Mixed Binary and Normal Annotations

The general formulation of Equation (1) allows different link functions $g_{jk}(\cdot)$ for different annotations, as well as different covariance structures $\boldsymbol{\Sigma}_j(\boldsymbol{\theta})$ to accommodate for correlations between the annotations (**Figure S1c**). In this section, we derive specific theoretical results for the EM algorithm when annotations are either conditionally bernoulli or normal random variables, i.e. all link functions $g_{jk}(\cdot)$ are either the identity or logistic link. We also introduce restrictions on the covariance matrices $\boldsymbol{\Sigma}_j(\boldsymbol{\theta})$ that allow for accurate approximations while greatly reducing the algorithm's computational cost. We call this algorithm MACIE for Multi-dimensional Annotation Class Integrative Estimation.

Suppose that conditionally on c_{ij} and $\boldsymbol{b}_{ij}$, the first $L_j^{(1)}$ of the L_j outcomes $\boldsymbol{y}_{ij}$ follow a bernoulli distribution and the remaining $L_j^{(2)} = L_j - L_j^{(1)}$ outcomes follow a normal distribution. That is, for $k = 1,2,\dots, L_j^{(1)}$, y_{ijk} has distribution

$$f_{jk}(y_{ijk}|c_{ij}, b_{ijk}) = \exp[y_{ijk}\eta_{ijk} - \log\{1 + \exp(\eta_{ijk})\}]$$

where $\mu_{ijk} = \exp(\eta_{ijk})/\{1 + \exp(\eta_{ijk})\}$ and $V_{ijk} = \mu_{ijk}(1 - \mu_{ijk})$. Then for $k = L_j^{(1)} + 1, L_j^{(1)} + 2, \dots, L_j$, y_{ijk} has the distribution

$$f_{jk}(y_{ijk}|c_{ij}, b_{ijk}) = \exp\left[\frac{\left\{y_{ijk}\mu_{ijk} - \frac{\mu_{ijk}^2}{2}\right\}}{\phi_{jk}} - \frac{1}{2}\left\{\frac{y_{ijk}^2}{\phi_{jk}} + \log(2\pi\phi_{jk})\right\}\right],$$

where $\mu_{ijk} = \eta_{ijk}$ and $V_{ijk} = \phi_{jk}$.

If $\boldsymbol{\Sigma}_j(\boldsymbol{\theta})$ is left unstructured, then the EM algorithm will need to estimate $L_j(L_j + 1)/2$ parameters for the covariance matrix of class j. An even greater computational challenge is that the multivariate Gauss-Hermite quadrature will require T^{L_j} fixed abscissas. Thus, to reduce the number of model parameters and to make the algorithm computationally feasible, we assume that $\boldsymbol{b}_{ij} = \boldsymbol{\Lambda}_j\boldsymbol{f}_{ij}$ where $\boldsymbol{f}_{ij}$ is an unobserved vector of length $P_j < L_j$ that follows $MVN(\boldsymbol{0}, \mathbf{I})$. Then for the E-step,

$$\int g(\boldsymbol{y}_{ij'}, \boldsymbol{c}_{ij'}, \boldsymbol{b}_{ij'})^{1(j=j')} f(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij}$$
$$= \int g(\boldsymbol{y}_{ij}, \boldsymbol{c}_{ij}, \boldsymbol{b}_{ij})^{1(j=j')} f(\boldsymbol{y}_{ij}|c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{f}_{ij}) \mathrm{d}\boldsymbol{f}_{ij},$$

so that integration is over a P_j-dimensional space as opposed to an L_j-dimensional space. The assumption $\boldsymbol{b}_{ij} = \boldsymbol{\Lambda}_j\boldsymbol{f}_{ij}$ forms the basis of factor analysis models[7] and is appropriate when the relationship between L_j manifest variables is thought to be primarily a result of the relationship between P_j underlying variables. For functional annotations, the underlying variables are likely to correspond to different approaches measuring the same element. As in factor analysis, the larger the factor loading λ_{jkp}, the more the jkth annotation is said to "load" on the pth factor.

For the L_{j1} binary outcomes, substituting the appropriate quantities into Equation (6) leads to the following expected score functions for variant i on outcome jk:

$$E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\beta}_{jk}) = \textstyle\sum_{c_{ij}=0}^{1} \left(\int \boldsymbol{x}_{ij}(y_{ijk} - \mu_{ijk}) \cdot f(\boldsymbol{b}_{ij} | \boldsymbol{y}_{ij}, c_{ij}) \mathrm{d}\boldsymbol{b}_{ij} \right) p(c_{ij} | \boldsymbol{y}_i), \qquad (10)$$

$$E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\Lambda}_{jk}) = \sum_{c_{ij}=0}^{1} \left(\int \boldsymbol{f}_{ij}(y_{ijk} - \mu_{ijk}) \cdot f(\boldsymbol{b}_{ij} | \boldsymbol{y}_{ij}, c_{ij}) \mathrm{d}\boldsymbol{b}_{ij} \right) p(c_{ij} | \boldsymbol{y}_i),$$

where $\boldsymbol{\Lambda}_{jk}$ is the kth column vector of $\boldsymbol{\Lambda}_j^T$.

To update estimates for $\boldsymbol{\beta}_{jk}$ using a one-step Fisher scoring algorithm, we consider a Taylor series expansion of the expected score function (Equation (10)) about the true parameter $\boldsymbol{\beta}_{jk}$,

$$E_{\boldsymbol{c},\boldsymbol{b}} S_i(\widehat{\boldsymbol{\beta}}_{jk}) \approx E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\beta}_{jk}) + \left\{ \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\beta}_{jk}) \right\} (\widehat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk}).$$

Since $E_{\boldsymbol{c},\boldsymbol{b}} S(\widehat{\boldsymbol{\beta}}_{jk}) = \sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}} S_i(\widehat{\boldsymbol{\beta}}_{jk}) = \boldsymbol{0}$, and assuming regularity conditions that allow the interchange of differentiation and integration, we have

$$E_{\boldsymbol{c},\boldsymbol{b}} S(\boldsymbol{\beta}_{jk}) \approx \left\{ \sum_{i=1}^{N} I_i(\boldsymbol{\beta}_{jk}) \right\} (\widehat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk}),$$

where I_i is the ith variant's contribution to the observed data Fisher information associated with the jkth outcome:

$$I_i(\boldsymbol{\beta}_{jk}) = -\frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\beta}_{jk}) = -\sum_{\boldsymbol{c}_i \in \{0,1\}^M} \left(\int \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} S_i(\boldsymbol{\beta}_{jk}) f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{c}_i) \mathrm{d}\boldsymbol{b}_i \right) p(\boldsymbol{c}_i | \boldsymbol{y}_i).$$

The expected information is obtained by taking an additional expectation with respect to the observed outcomes $\boldsymbol{y}_i$:

$$J_i(\boldsymbol{\beta}_{jk}) = -E_{\boldsymbol{y}_i} \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} E_{\boldsymbol{c},\boldsymbol{b}} S_i(\boldsymbol{\beta}_{jk}).$$

Interchanging derivatives and expectations yields

$$J_i(\boldsymbol{\beta}_{jk}) = -\sum_{\boldsymbol{c}_i \in \{0,1\}^M} \left(\int E_{\boldsymbol{y}_i} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}_{jk}^T} S_i(\boldsymbol{\beta}_{jk}) \right\} f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{c}_i) \mathrm{d}\boldsymbol{b}_i \right) p(\boldsymbol{c}_i | \boldsymbol{y}_i).$$

For binary outcomes with logistic link, the expected information is

$$J_i(\boldsymbol{\beta}_{jk}) = \sum_{\boldsymbol{c}_i\in\{0,1\}^M}\left(\int \boldsymbol{x}_{ij}\mu_{ijk}(1-\mu_{ijk})\boldsymbol{x}_{ij}^T f(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{c}_i)\mathrm{d}\boldsymbol{b}_i\right)p(\boldsymbol{c}_i|\boldsymbol{y}_i). \qquad (11)$$

Equations (10) and (11) yield the following scoring algorithm at iteration $r+1$:

$$\widehat{\boldsymbol{\beta}}_{jk}^{(r+1)} = \widehat{\boldsymbol{\beta}}_{jk}^{(r)} + \left(\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{x}_{ij}\hat{\mu}_{ijk}^{(r)}\left(1-\hat{\mu}_{ijk}^{(r)}\right)\boldsymbol{x}_{ij}^T\right]\right)^{-1}\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{x}_{ij}\left(y_{ijk}-\hat{\mu}_{ijk}^{(r)}\right)\right]. \qquad (12)$$

Similarly,

$$\widehat{\boldsymbol{\Lambda}}_{jk}^{(r+1)} = \widehat{\boldsymbol{\Lambda}}_{jk}^{(r)} + \left(\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{f}_{ij}\hat{\mu}_{ijk}^{(r)}\left(1-\hat{\mu}_{ijk}^{(r)}\right)\boldsymbol{f}_{ij}^T\right]\right)^{-1}\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{f}_{ij}\left(y_{ijk}-\hat{\mu}_{ijk}^{(r)}\right)\right]. \qquad (13)$$

For the $L_j^{(2)}$ normal outcomes, contributions to the complete data score functions for each variant i are

$$S_i(\boldsymbol{\beta}_{jk}) = \frac{1}{\phi_{jk}}\boldsymbol{x}_{ij}e_{ijk},$$

$$S_i(\boldsymbol{\Lambda}_{jk}) = \frac{1}{\phi_{jk}}\boldsymbol{f}_{ij}e_{ijk},$$

$$S_i(\phi_{jk}) = -\frac{1}{2\phi_{jk}} + \frac{1}{2\phi_{jk}^2}e_{ijk}^2,$$

where $e_{ijk} = y_{ijk} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{jk} - \boldsymbol{f}_{ij}^T\boldsymbol{\Lambda}_{jk}$. It follows that

$$\widehat{\boldsymbol{\beta}}_{jk}^{(r+1)} = \left(\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\right]\right)^{-1}\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{x}_{ij}\left(y_{ijk}-\boldsymbol{f}_{ij}^T\widehat{\boldsymbol{\Lambda}}_{jk}^{(r)}\right)\right], \qquad (14)$$

$$\widehat{\boldsymbol{\Lambda}}_{jk}^{(r+1)} = \left(\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{f}_{ij}\boldsymbol{f}_{ij}^T\right]\right)^{-1}\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\boldsymbol{f}_{ij}\left(y_{ijk}-\boldsymbol{x}_{ij}^T\widehat{\boldsymbol{\beta}}_{jk}^{(r)}\right)\right], \qquad (15)$$

$$\hat{\phi}_{jk}^{(r+1)} = \frac{1}{N}\sum_{i=1}^{N} E_{\boldsymbol{c},\boldsymbol{b}}\left[\hat{e}_{ijk}^{(r)2}\right]. \qquad (16)$$

Beginning with reasonable initial estimates of the parameters, MACIE proceeds by first using the E-step to obtain the desired expectations relative to the posterior distribution. Given those estimates, MACIE then solves the expected score equations to obtain new parameter estimates or one-step updates according to Equations (5), (12)-(16). The algorithm proceeds until the relative changes in all estimated parameters are sufficiently small ($< 10^{-4}$) with a maximum of 200 iterations.

Data Analysis Using the MACIE GLMM

We used the proposed framework to fit the MACIE GLMM models for (1) non-synonymous coding variants and (2) non-coding and synonymous coding variants separately. For non-synonymous coding variants, we considered fitting a two-class MACIE model ($M = 2$) where the damaging protein function class included 4 protein substitution scores: SIFT, PolyPhenDiv, PolyPhenVar, and Mutation Assessor, with two latent factors of $\mathbf{\Sigma}_1$; and the evolutionary conserved class included 8 conservation scores: GERP_NR, GERP_RS, PhyloPri, PhyloPla, PhyloVer, PhastPri, PhastPla, and PhastVer, with two latent factors of $\mathbf{\Sigma}_2$ (**Table S1**). Thus, the MACIE score predicted for each non-synonymous coding variant is a vector of length 4, representing the estimated joint posterior probabilities of belonging to (0,1) - "not damaging protein functional and conserved"; (1,0) - "damaging protein functional and not conserved"; (0,0) - "not damaging protein functional and not conserved"; (1,1) - "both damaging protein functional and conserved". The MACIE GLMM regression parameter estimates from the training set of non-synonymous coding variants are presented in **Table S2**.

For non-coding and synonymous coding variants, we considered fitting a two-class MACIE model ($M = 2$), where the evolutionary conserved class included the same 8 conservation scorers as the non-synonymous coding model, with two latent factors of $\mathbf{\Sigma}_1$, and the regulatory class included a total of 28 transformed epigenetic scores scores, consisting of 3 histone marks and 12 open chromatin marks from the ENCODE Project, 3 transcription factor binding site scores, GC content, CpG content, 5 chromatin state probabilities derived from the 15 state ChromHMM model, a background selection score, and 2 physical distance metrics, with three latent factors of $\mathbf{\Sigma}_2$ (**Table S1**). Detailed information on pre-processing steps for the epigenetic scores are given in **Table S3**. Thus, the MACIE score predicted for each non-coding or synonymous coding variant is also a vector of length 4, representing the estimated joint posterior probabilities of belonging to (0,1) - "not conserved and regulatory functional"; (1,0) - "conserved and not regulatory functional"; (0,0) - "not conserved and not regulatory functional"; (1,1) - "both conserved and regulatory functional". The MACIE GLMM regression parameter estimates from the training set of non-coding and synonymous coding variants are presented in **Table S4**.

Heritability Enrichment Analysis Using Stratified LD Score Regression

Following Li et al.,[8] we performed a stratified LD score regression (S-LDSC) heritability enrichment analysis for the 2 lipid traits from the ENGAGE Consortium,[9] LDL-C and HDL-C, using 8 functional annotations, namely fitCons, CADD, LINSIGHT, FATHMM-XF, EIGEN, DANN, GenoCanyon, and MACIE-anyclass. Specifically, we considered the common SNVs (MAF > 5%) in the International HapMap Project Phase 3 (HapMap3),[10] as in the previous work.[11; 12] For each of the 8 functional annotations, following Li et al., we created a binary functional category by labeling the top 30% SNVs as 1 (meaning these SNVs are functional; same as Li et al.[8]) and evaluated the heritability enrichment of the binary annotation using S-LDSC conditional on the set of 52 baseline annotations;[12] the heritability enrichment is defined as the ratio of the proportion of heritability explained by the annotated variants in the functional category and the proportion of variants in the functional category.[8] The results showed that MACIE-anyclass, CADD, and DANN showed significant enrichment for both lipid traits; FitCons, LINSIGHT, EIGEN, and GenoCanyon showed significant enrichment only for HDL-C; while FATHMM-XF showed non-significant enrichment (**Figure S8**, **Table S16**). This analysis suggests that the MACIE score is informative for prioritizing GWAS trait-associated variants.

References

1. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336-2337.
2. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological), 1-38.
3. Sammel, M.D., Ryan, L.M., and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59, 667-678.
4. Little, R.J., and Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley, 1987.
5. McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models, Second Edition.(Taylor & Francis).
6. Abramowitz, M., and Stegun, I.A. (1964). Handbook of mathematical functions: with formulas, graphs, and mathematical tables.(Courier Corporation).
7. Lawley, D., and Maxwell, A. (1962). Factor analysis as a statistical method. Journal of the Royal Statistical Society Series D (The Statistician) 12, 209-229.
8. Li, B., Lu, Q., and Zhao, H. (2019). An evaluation of noncoding genome annotation tools through enrichment analysis of 15 genome-wide association studies. Briefings in Bioinformatics 20, 995-1003.
9. Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., and Timonen, S. (2015). The impact of low-frequency and rare variants on lipid levels. Nature genetics 47, 589-597.
10. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.
11. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., and Schizophrenia Working Group of the Psychiatric Genomics, C. (2015). LD Score regression distinguishes confounding

from polygenicity in genome-wide association studies. Nature Genetics 47, 291-295.

12. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., and Farh, K. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics 47, 1228.