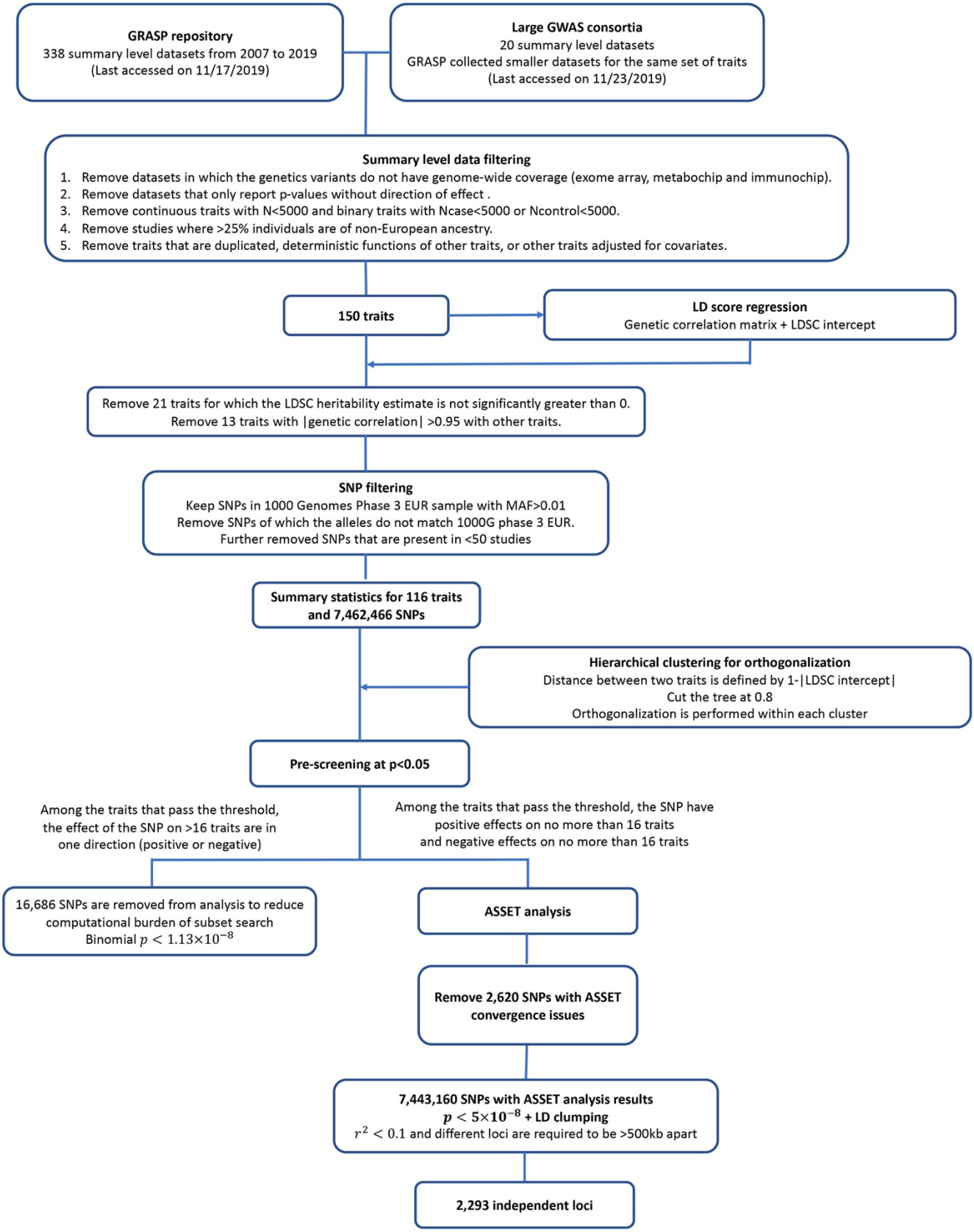
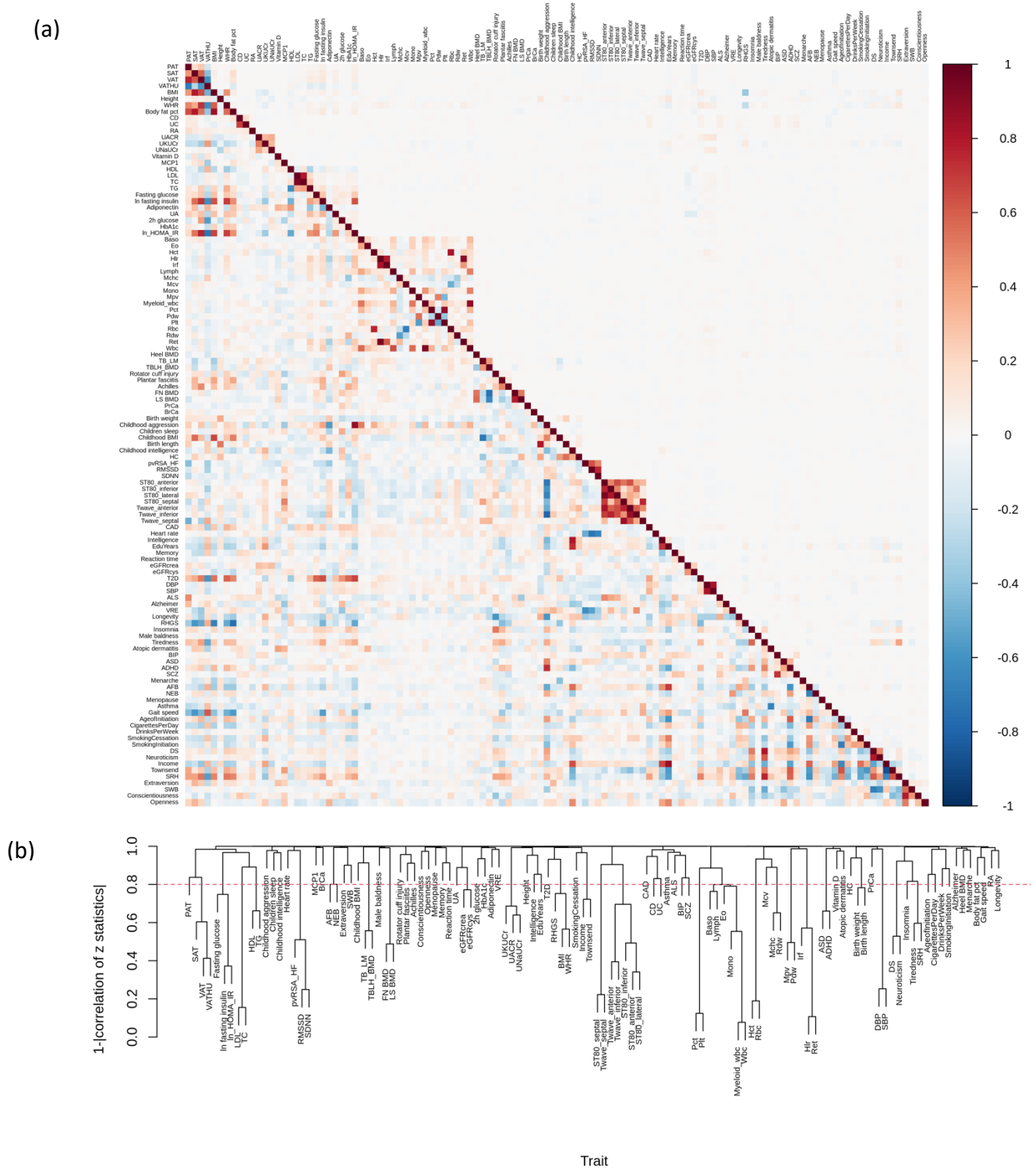
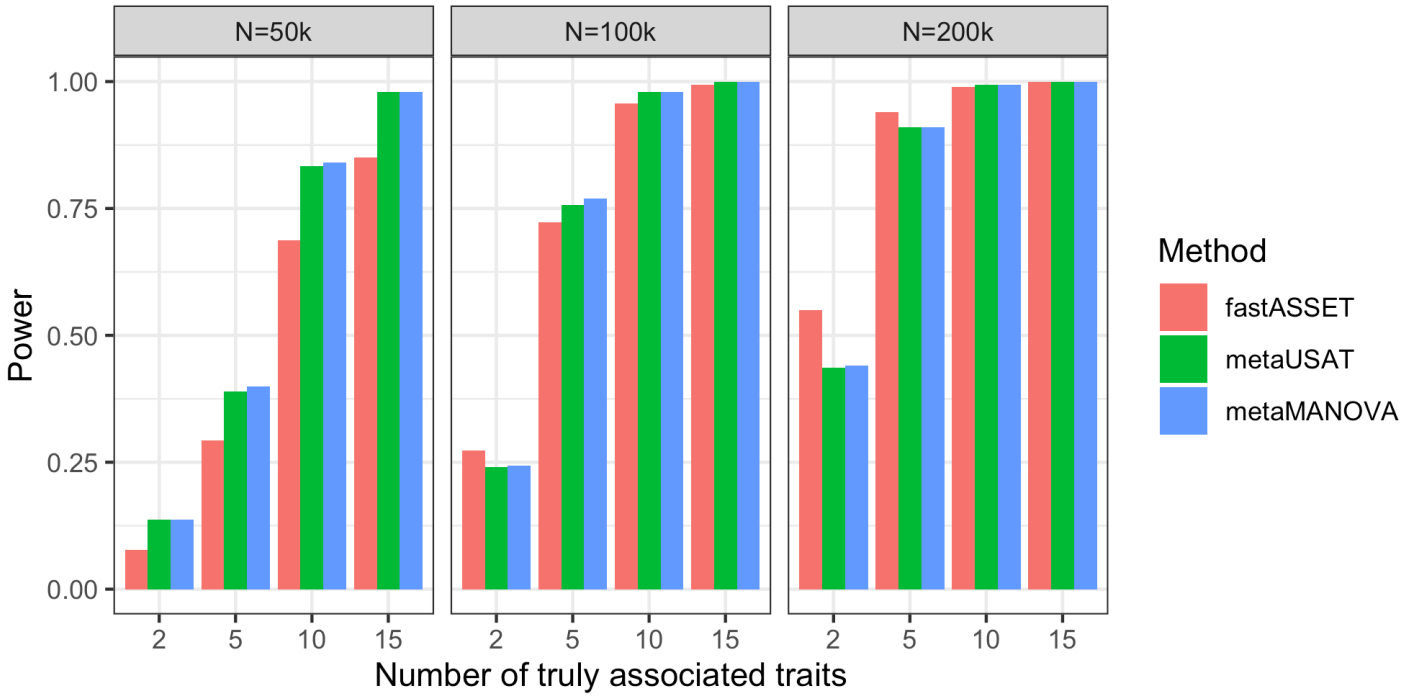
Supplementary Figures



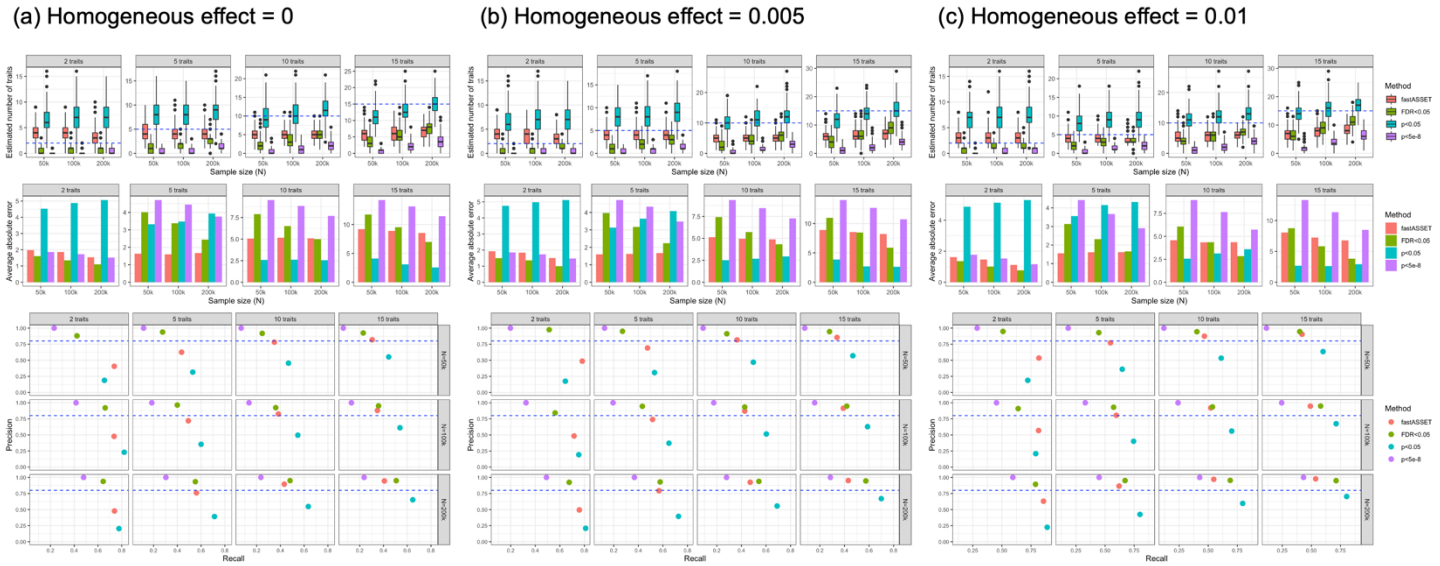
Supplementary Figure 1. Data preprocessing and analysis pipeline. Details are provided in Methods.



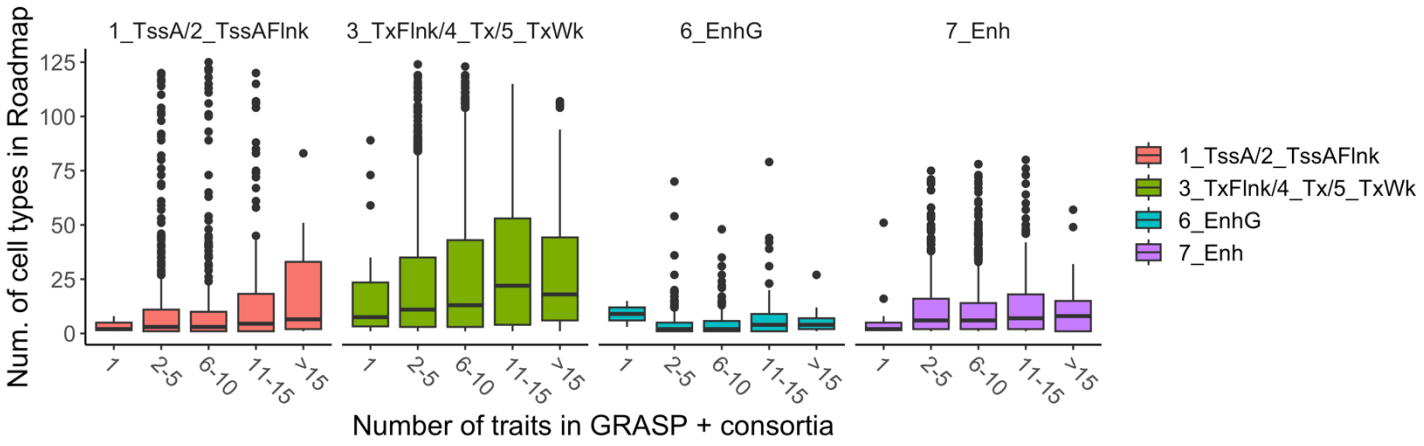
Supplementary Figure 2. Genetic correlation and sample overlap across traits. (a) Genetic correlation (lower diagonal) across 116 traits estimated by LD score regression (LDSC) and correlation of z statistics under the global null hypothesis (upper diagonal, equal to LDSC intercept, representing sample overlap). Pairs of studies with substantially nonzero LDSC intercept needs to be de-correlated before ASSET analysis (Supplementary Notes). De-correlation is performed within clusters of traits defined by (b) hierarchical clustering based on LDSC intercept.



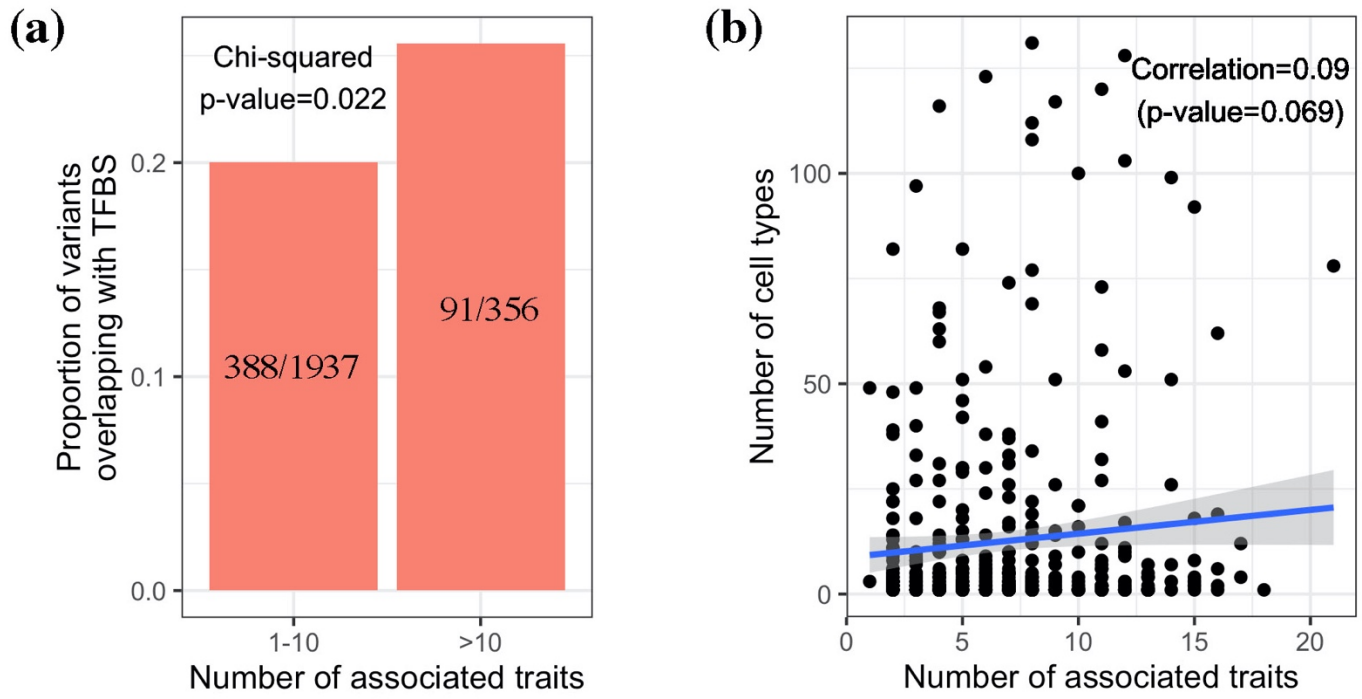
Supplementary Figure 3. Power of fastASSET for detecting global association in simulation studies. The null hypothesis is that the SNP is not associated with any of the trait under consideration. GWAS regression coefficients for one SNP j and 116 traits ($\hat{\beta}_j$, vector of length 116) are simulated from model $\hat{\beta}_j = \beta_j + e_j$, where e_j is the error term generated from multivariate normal distribution reflecting realistic correlation across traits. True effect β_j is simulated by first randomly selecting a set of K=2, 5, 10 or 15 traits (columns) which have true associations with the SNP, followed by generating the effect size from $\beta_{fix} + N(\text{mean} = 0, \text{variance} = 0.01^2)$. Around half of the K traits (round(K/2)) have $\beta_{fix} = -0.01$ and the rest have $\beta_{fix} = 0.01$, which reflects bidirectional fixed effects. Effect size heterogeneity is reflected by $N(0, 0.01^2)$. We vary the sample size N, assumed to be the same across all traits, to 50k, 100k and 200k (rows). Power is evaluated at genome-wide significance threshold $p < 5 \times 10^{-8}$. This simulation procedure is repeated 300 times for each setting. The power of fastASSET is compared to that of metaUSAT and metaMANOVA. See Supplementary Notes for details of the simulation settings.



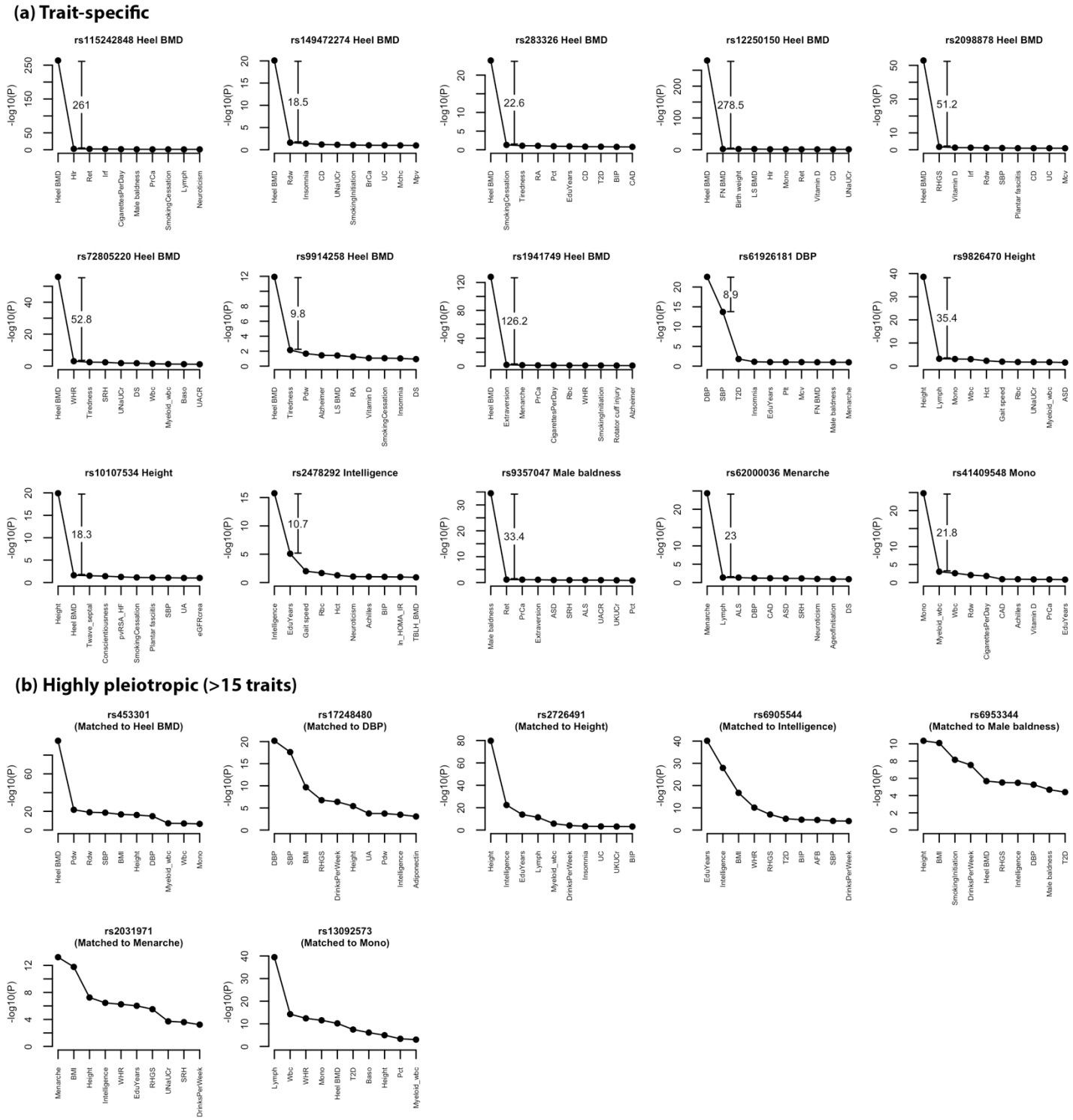
Supplementary Figure 4. Additional Simulation studies with heterogeneous effect size and genetic correlation. GWAS regression coefficients for 116 traits are simulated from model $\hat{\beta}_j = \beta_j + e_j$ (see legend of Supplementary Figure 3 for details). True effect β_j is simulated by first randomly selecting a set of K=2, 5, 10 or 15 traits (columns) which have true associations with the SNP, followed by generating the effect size from $\beta_j = \text{Bernoulli}(0.5) \times \beta_{hom} + \beta_{het}$. The homogeneous effects β_{hom} can take three values: 0, 0.005, and 0.01 in different simulation settings. The Bernoulli(0.5) distribution allows the effect to be in both directions. Heterogeneous effects are simulated from multivariate normal distribution $\beta_{het} \sim N(\mathbf{0}, \frac{\Sigma_g^{(K)}}{2000})$. Here $\Sigma_g^{(K)}$ is a $K \times K$ submatrix of the genetic covariance matrix in our European ancestry dataset estimated from LD score regression. We scale $\Sigma_g^{(K)}$ by 2,000 to emulate the scenarios where 2,000 causal SNPs contribute to heritability and genetic correlation. Average sample size across traits (N) varies among 50k, 100k and 200k. Trait-specific sample size is randomly generated from Uniform[0.5N, 1.5N]. This simulation is repeated 300 times for each setting. (a) $\beta_{hom} = 0$. (b) $\beta_{hom} = 0.005$. (c) $\beta_{hom} = 0.01$.



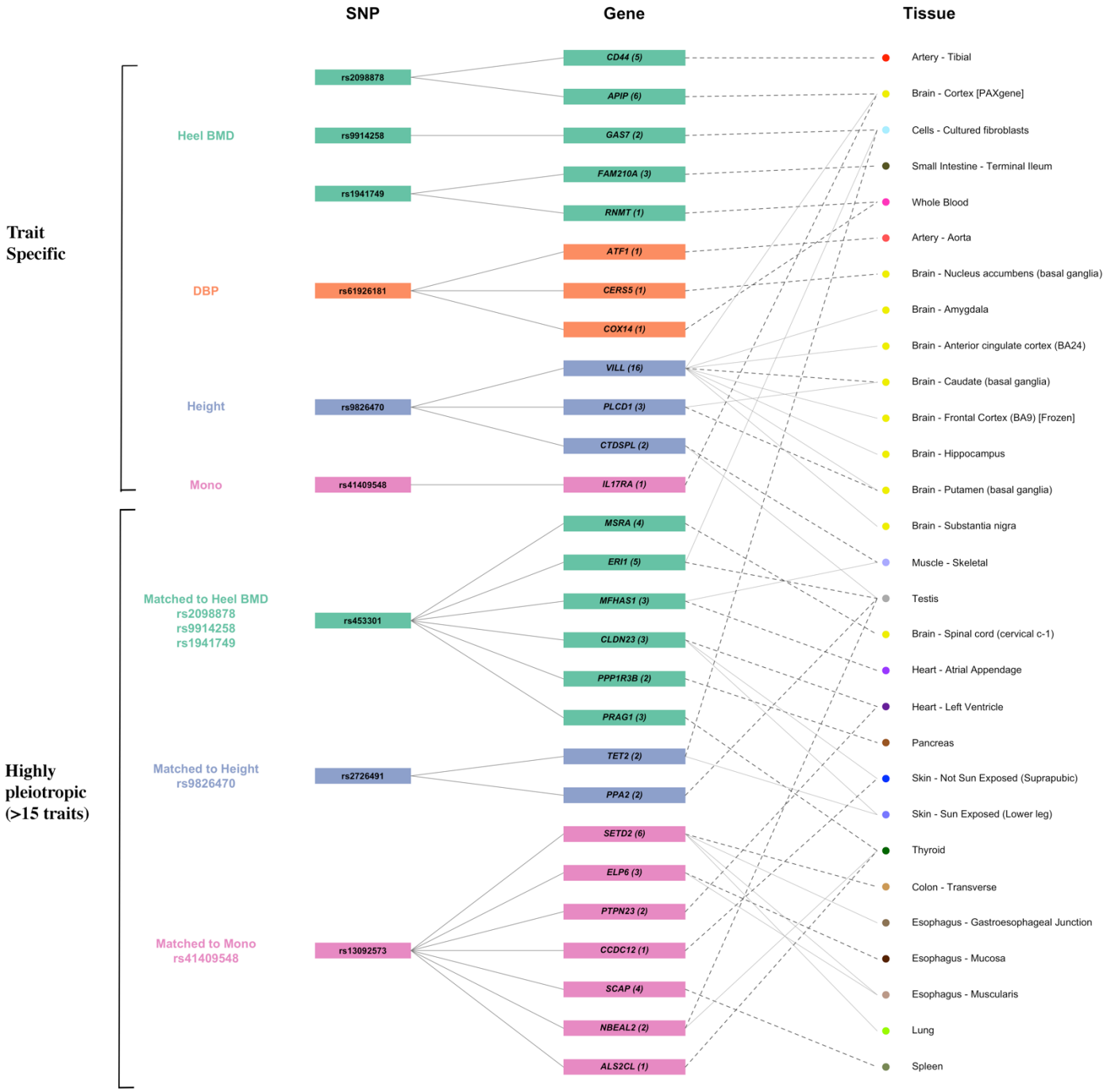
Supplementary Figure 5. Relationship between pleiotropy and number of cell types where the SNP is in active chromatin state learned by ChromHMM. Chromatin states are learned from Roadmap Epigenomics Consortium data. Boxplots show the median (centerline) and first and third quartiles (lower and upper hinges) of the distribution.



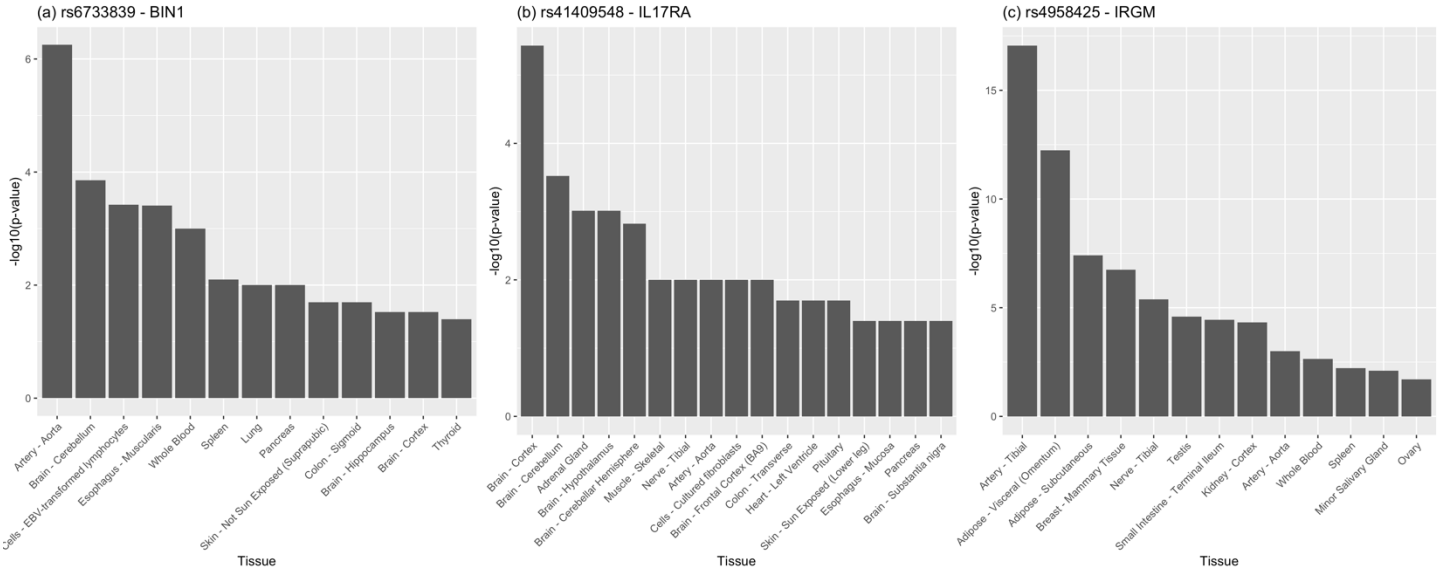
Supplementary Figure 6: Relationship between pleiotropy and additional functional annotations. (a) Overlap with transcription factor binding sites (TFBS) from the JASPAR and HOCOMOCO databases. (b) Cell type specificity of enhancer-gene connection identified by the activity-by-contact (ABC) model; the y axis is the number of cell types for which the enhancer overlapping with the SNP affects at least one target gene by the ABC model. After adjusting for other annotations (LD score, number of tissues for which the variant is an eQTL, number of eGenes, number of cell types for which the variant is in active chromatin state), the association in (a) remains significant (p-value=0.021), while the association in (b) disappears (partial correlation=0.02, p-value=0.65). See Methods for details.



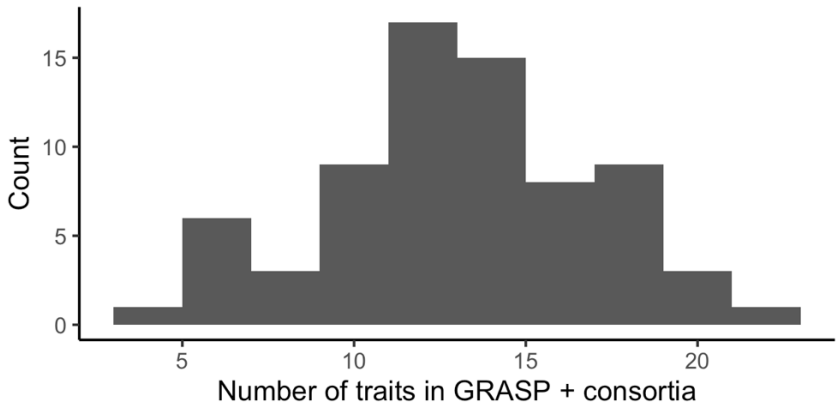
Supplementary Figure 7. Top 10 associations of trait-specific SNPs for quantitative traits and matched highly pleiotropic SNPs. (a) Trait-specific SNPs for quantitative traits. (b) Matched highly pleiotropic SNPs. Traits are ordered by descending order or -log10(p-value) in each panel. Each trait-specific SNP is matched to the highly pleiotropic SNP (>15 traits) that has the smallest association p-value with the trait (matched trait shown in the parentheses). See **Methods** for details of the matching procedure.



Supplementary Figure 8. Cis-regulatory effects of trait-specific SNPs for quantitative traits and matched highly pleiotropic SNPs. For each SNP, we show its protein-coding eGenes (q-value < 0.05) and the corresponding top tissues in GTEx v8. The top tissues for each variant-gene pair are defined as those with eQTL effect size >0.7*(the largest effect size among significant tissues for this variant-gene pair); the tissue harboring the largest effect is highlighted by darker dashed lines. In addition, we annotate each gene name by the total number of associated tissues (q-value < 0.05) regardless of effect size. Each trait-specific SNP is matched to the highly pleiotropic SNP (>15 traits) that have the strongest association with the trait (by p-value). See **Methods** for details of the matching procedure. The pleiotropic SNP matched to diastolic blood pressure (DBP) is not an eQTL for any gene-tissue pair and hence omitted. Pseudogenes and non-coding RNAs are excluded. BMD: bone mineral density; mono: monocyte count.



Supplementary Figure 9. Association between three trait-specific SNPs and an eGene in multiple tissues. (a) rs6733839 is a trait-specific SNP for Alzheimer's disease. (b) rs41409548 is a trait-specific SNP for monocyte count. (c) rs4958425 is a trait-specific SNP for Crohn's disease. P-values are nominal p-values derived from linear regression (two-sided t-test for regression slope). For each SNP-gene pair, only tissues with p-value <0.05 are shown.



Supplementary Figure 10. Distribution of level of pleiotropy for the SNPs that are removed from ASSET analysis. These SNPs represent 74 independent loci with $r^2 < 0.1$ and >500kb apart.

Supplementary Notes

1 Details of the data filtering steps

We collect 358 summary-level datasets that cover a wide range of phenotypes, including 338 datasets from NIH GRASP repository and 20 from large GWAS consortia. We apply the following filtering steps to the datasets:

1. Remove datasets for which important information is missing. Scenarios include: 1) The genotypes are collected by a special chip without genome-wide coverage (metabochip, immunochip, exome array, etc). 2) The data only provide p-values for associations, but do not provide the direction of SNP effect on the trait. 3) The association testing was performed using nonstandard approaches or stratified analysis based on covariates. 4) The README file is missing hence the meaning of each column is unclear.

2. Remove studies with small sample size. For continuous traits, we remove studies with sample size >5,000; for binary traits, we remove studies with <5,000 cases or <5,000 controls.

3. Remove studies where >25% individuals are of non-European ancestry.

4. Remove duplicated traits. Since the GRASP repository includes studies spanning over a decade, there are often multiple studies for the same trait (e.g., multiple versions of height and BMI GWASs from the GIANT Consortium). Among the studies for the same trait, we keep the study with the largest sample size and discard the rest of them.

2 ASSET

2.1 ASSET for independent samples across traits

The original ASSET is built on top of meta-analysis across different phenotypes (Bhattacharjee et al, 2012). Let $(\beta_k, s_k),\ k = 1, 2, ..., K$ denote the estimates of regression parameter and its standard error for a given SNP and K traits. Let $z_k = \beta_k / s_k$ be the corresponding z-statistics. The standard meta-analysis uses the following statistic to test for association

$$z_{meta} = \sum_{k=1}^{K} w_k z_k$$

where $w_k = \frac{\sqrt{n_k}}{\sqrt{\sum_{k=1}^{K} n_k}}$. For continuous traits, $s_k = 1/\sqrt{n_k}$ when both the genotype and the K traits are standardize to have unit variance. For case-control binary traits, $s_k = 1/\sqrt{n_k}$ when the genotype is standardized and $n_k = \frac{n_{case} n_{control}}{n_{case}+n_{control}}$

Since it is unknown which traits are truly associated with the SNP, including traits that are not associated can lead to loss of power. Therefore, the ASSET method conducts a search through all subsets of traits and and finds the subset that gives maximum meta-analysis z-statistic. Denote the meta-analysis z-statistic based on subset S by $Z(S)$, which can be expressed as

$$Z(S) = \sum_{k \in S} w_k z_k$$

This maximum meta-analysis z-statistic is used as the new test statistic:

$$Z_{max-meta} = max_{S \in \boldsymbol{S}} |Z(S)|$$

$\boldsymbol{S}$ indexes all possible subsets involved in the maximum. The p-value is approximated using discrete local maxima.

This test statistic is powerful when the effect of the SNP on all the traits are in the same direction, which is unlikely in our analysis due to the large number of traits. Hence we adopt the two-side approach:

1. Conduct subset search within the positive and negative z statistics separately.
2. Compute the conditional p-values given the signs of the z-statistics, denoted by $\tilde{P}^{+}_{DLM}$ and $\tilde{P}^{-}_{DLM}$.
3. Combine the p-values using Fisher's method

$$Z^{(2)}_{max-meta} = -2[\log \tilde{P}^{+}_{DLM} + \log \tilde{P}^{-}_{DLM}].$$

The new statistic $Z^{(2)}_{max-meta}$ follows a chi-squared distribution.

2.2 ASSET with sample overlap across traits

Due to substantial sample overlap across traits in our study, we adopt the version of ASSET that allows for correlation across z-statistics under the null hypothesis (denoted by $\boldsymbol{\rho}^{(z)}$). For subset S, denote the column vector of z-statistics by $\boldsymbol{z}_S$, the column vector of weights by $\boldsymbol{w}_S = (w_k)_{k \in S} \propto (\sqrt{n_k})_{k \in S}$, and the correlation matrix of z-statistics by $\boldsymbol{\rho}^{(z)}_S$. The meta-analysis z-statistic for subset S is

$$Z(S) = \frac{\boldsymbol{w}_S^T \boldsymbol{z}_S}{\sqrt{\boldsymbol{w}_S^T \boldsymbol{\rho}^{(z)}_S \boldsymbol{w}_S}}.$$

The discrete local maxima approximation of p-value is

$$P_{DLM} = \sum_{s \in \boldsymbol{S}} \int_T^{\infty} 2 \prod_{k=1}^{K} P(|Z_{s,\pm k}| < z | Z_s = z) \phi(z) dz$$

Here $Z_{s,\pm k}$ is the k-th neighbor of the current subset s obtained by adding the k-th study if it is not already included and dropping it otherwise. Conditional probability $P(|Z_{s,\pm k}| < z | Z_s = z)$ can be computed from multivariate normal distribution

$$\boldsymbol{z} = (z_1, ..., z_K)^T \sim N(\mathbf{0}, \boldsymbol{\rho}^{(z)}).$$

The other steps are identical to ASSET for independent samples across studies (Section 2.1). When $\boldsymbol{\rho}^{(z)} = I$, it reduces to ASSET for independent samples.

3 Simulation studies

3.1 Null simulations

We conduct simulation studies to evaluate the type I error control of fastASSET. We showed in our previous publications that simulations models based on individual genotypes implies a model based on summary statistics (Qi and Chatterjee, 2018; Qi and Chatterjee, 2021). In particular, a null model implies that the summary statistics follow a multivariate normal distribution centered at 0. Hence we simulate the summary statistics directly as follows:

$$\hat{\beta}_j \sim N(\mathbf{0}, \Sigma/N). \tag{1}$$

Here $\hat{\beta}_j$ is a vector of GWAS regression coefficients for a single SNP across multiple traits. We set sample size $N = 20,000$ for all the traits, and hence the GWAS standard error is $1/\sqrt{N}$. Here Σ is the covariance matrix of z-statistics under the null, incorporating both sample overlap and phenotypic correlation. Element (k, l) of this matrix is $\frac{\rho_{kl} N_{kl}}{N}$, where ρ_{kl} is the correlation between traits k and l and N_{kl} is the number of overlapping samples. To obtain a realistic covariance structure, we set Σ as the LD score regression intercept matrix for the 116 traits in our study. We apply fastASSET to the simulated summary statistics. Since we did not simulate LD structure, it is not possible to use LD score regression to estimate the $\boldsymbol{\rho}^{(z)}$ matrices. Instead, we directly use Σ in place of $\boldsymbol{\rho}^{(z)}$ and treat it as known. Here we conduct 20 million independent simulations ($j = 1, 2, ..., 20,000,000$) to evaluate the type I error rate.

3.2 Non-null simulations for power evaluation

To evaluate the power of fastASSET, we simulate the summary statistics as a sum of two vectors of length 116:

$$\hat{\beta}_j = \beta_j + e_j, \quad j = 1, 2, ..., 300$$

where β_j is the true effect and e_j is the error. The error vector is simulated from $e_j \sim N(\mathbf{0}, \Sigma/N)$, the same distribution used in the null simulations. We still use the same sample size N for all traits, but vary it to 20k, 50k, 100k and 200k across different simulations.

To simulate the vector of true effects β_j, we randomly select K traits (denote the indices of selected traits by $\mathcal{K}$) that have true association with the SNP. We vary K to 2, 5, 10 or 15. We simulate the k-th element of β_j by $\beta_{jk} \sim \beta_{fix} + N(0, 0.01^2)$ if $k \in \mathcal{K}$, and $\beta_{jk} = 0$ otherwise. Around half of the K traits ($round(K/2)$) have $\beta_{fix} = -0.01$ and the rest have $\beta_{fix} = 0.01$, which reflects bidirectional fixed effects. The normal distribution $N(0, 0.01^2)$ allows heterogeneous effect size across traits. Note that we only simulate 300 SNPs, which is sufficient for the evaluation of power and estimation of pleiotropy.

For both type I error and power, we compare fastASSET to metaUSAT and metaMANOVA (Ray and Boehnke, 2018), two statistical methods for multi-trait association testing.

Supplementary references

1. Bhattacharjee, Samsiddhi, et al. "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits." The American Journal of Human Genetics 90.5 (2012): 821-835.

2. Bulik-Sullivan, Brendan, et al. "An atlas of genetic correlations across human diseases and traits." Nature genetics 47.11 (2015): 1236-1241.

3. Qi, Guanghao, and Nilanjan Chatterjee. "Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits." PLoS genetics 14.10 (2018): e1007549.

4. Qi, Guanghao, and Nilanjan Chatterjee. "A comprehensive evaluation of methods for Mendelian randomization using realistic simulations and an analysis of 38 biomarkers for risk of type 2 diabetes." International Journal of Epidemiology (2021).

5. Ray, Debashree, and Michael Boehnke. "Methods for meta-analysis of multiple traits using GWAS summary statistics." Genetic epidemiology 42.2 (2018): 134-145.