A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics

Authors

Zixuan Zhang, Junghyun Jung, Artem Kim,
Noah Suboc, Steven Gazal, Nicholas Mancuso

Correspondence

zzhang39@usc.edu (Z.Z.),
nmancuso@usc.edu (N.M.)

By leveraging numerous trait-associated genetic variants identified in genome-wide association studies (GWASs) across thousands of traits, Zhang et al. propose FactorGo, a scalable factor analysis model, to identify and characterize pleiotropic components using biobank GWAS summary data. The authors demonstrate that FactorGo improves our biological understanding of shared etiologies across thousands of GWASs.

Zhang et al., 2023, The American Journal of Human Genetics *110*, 1863–1874
November 2, 2023 © 2023 American Society of Human Genetics.
https://doi.org/10.1016/j.ajhg.2023.09.015

A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics

Zixuan Zhang,[1,*] Junghyun Jung,[1] Artem Kim,[1] Noah Suboc,[1] Steven Gazal,[1,2,3,4] and Nicholas Mancuso[1,2,3,4,*]

Summary

Genome-wide association studies (GWASs) across thousands of traits have revealed the pervasive pleiotropy of trait-associated genetic variants. While methods have been proposed to characterize pleiotropic components across groups of phenotypes, scaling these approaches to ultra-large-scale biobanks has been challenging. Here, we propose FactorGo, a scalable variational factor analysis model to identify and characterize pleiotropic components using biobank GWAS summary data. In extensive simulations, we observe that FactorGo outperforms the state-of-the-art (model-free) approach tSVD in capturing latent pleiotropic factors across phenotypes while maintaining a similar computational cost. We apply FactorGo to estimate 100 latent pleiotropic factors from GWAS summary data of 2,483 phenotypes measured in European-ancestry Pan-UK BioBank individuals ($N = 420{,}531$). Next, we find that factors from FactorGo are more enriched with relevant tissue-specific annotations than those identified by tSVD ($p = 2.58E{-}10$) and validate our approach by recapitulating brain-specific enrichment for BMI and the height-related connection between reproductive system and muscular-skeletal growth. Finally, our analyses suggest shared etiologies between rheumatoid arthritis and periodontal condition in addition to alkaline phosphatase as a candidate prognostic biomarker for prostate cancer. Overall, FactorGo improves our biological understanding of shared etiologies across thousands of GWASs.

Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic variants that associate with complex traits and diseases affecting multiple traits.[1–3] Investigating this pervasive pleiotropy has enabled elucidating broader biological mechanisms, identifying comorbidity due to genetic susceptibility, and discovering or repurposing of therapeutic targets.[4–6]

Previous works have proposed methods to identify pleiotropic components under two related, but distinct, camps of approaches. The first camp is to apply matrix factorization techniques (e.g., truncated singular value decomposition [tSVD]) on a matrix of GWAS summary data.[7–9] While matrix factorization provides a computationally efficient means of capturing apparent pleiotropic components, its model-free approach leaves unclear what parameters are inferred from noisy observations (in this case, effect-size estimates). The second camp of approaches is based on statistical models for genetic effects but is limited to the analysis of a small number of traits due to computational demands.[10–12] As more GWAS summary data become available in large biobanks,[13–15] it is important to develop a scalable model-based approach that allows exploring the phenome-wide shared genetic architecture, either known or unknown, to be genetically related *a priori*. Classical factor analysis provides an analogous approach toward summarizing shared latent factors in data; however, inference in high-dimensional biobank settings is computationally demanding, thus limiting the scope of applied analysis.

Here, to identify latent pleiotropic components across thousands of phenotypes, we present FactorGo, a factor analysis model on genetic associations using GWAS summary data. FactorGo models the uncertainty in genetic effect estimates and leverages an automatic relevance determination (ARD) prior to prune uninformative factors using a scalable variational Bayesian framework. Under extensive simulations, we find that FactorGo outperforms tSVD in reconstructing trait factor scores and is robust to model misspecifications. By analyzing thousands of phenotypes in Pan-UK Biobank (Pan-UKB), we identify alkaline phosphatase (ALP) as a candidate prognostic biomarker for prostate cancer (PCa). Moreover, we recapitulate previously reported brain-specific enrichment for BMI and reproductive system and muscular-skeletal enrichment for height. For disease traits, we learn the shared bacterial etiology between rheumatoid arthritis (RA) and periodontal condition. Taken together, our results demonstrate that FactorGo prioritizes biologically meaningful latent pleiotropic factors, which reflect shared biological mechanisms across traits.

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA; [3]Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[4]These authors contributed equally
*Correspondence: zzhang39@usc.edu (Z.Z.), nmancuso@usc.edu (N.M.)
https://doi.org/10.1016/j.ajhg.2023.09.015.

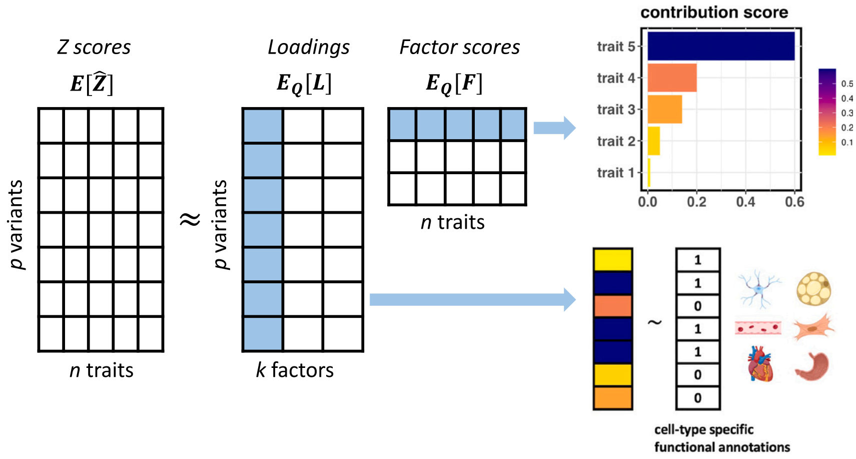


Figure 1. Overview of FactorGo
FactorGo decomposes the observed Z-score summary statistics of p variants in n traits to k pleiotropic factors.
The column vector of L is variant loadings and row vector of F is the trait factor score for each inferred factor as highlighted in light blue. Here, we plotted for $n = 5$, $p = 7$, and $k = 3$ for illustrative purposes. To identify traits characterizing a given factor, we calculated contribution scores of this factor across all traits (top arrow). To understand the biological function of a given factor, we regressed transformed variant loadings on cell-type-specific annotations using LD score regression (bottom arrow). The colors on transformed scores represent the magnitude of values.

Material and methods

FactorGo model

Here, we briefly describe the FactorGo generative model of observed GWAS summary data assuming correlations in effects arising due to pleiotropy. For a full account, please see details in Note S1. Briefly, FactorGo assumes observed Z scores (i.e., Z_i) are sampled around the scaled true genetic effects (i.e., $\sqrt{N_i}\beta_i$), which are decomposed into latent pleiotropic factors (i.e., $\beta_i = Lf_i + \mu$; see Figure 1). Formally, we model Z scores at p independent variants from the i^{th} GWAS Z_i as a linear combination of k shared latent variant loadings $L \in R^{p \times k}$ with trait-specific latent factor scores $f_i \in R^{k \times 1}$ and sampling variability ϵ_i as

$$Z_i = \sqrt{N_i}\beta_i + \epsilon_i = \sqrt{N_i}(Lf_i + \mu) + \epsilon_i = \sqrt{N_i}\left(\sum_k L_k f_{ik} + \mu\right) + \epsilon_i,$$

where N_i is the sample size for the i^{th} GWAS, μ is an intercept that reflects a global mean effect size across studies, and $\epsilon_i \sim N(0, \tau^{-1} I_p)$ reflects sampling variability around the estimate with residual heterogeneity across studies as precision scalar τ. In genome-wide data, we expect nearby summary statistics to be correlated due to linkage disequilibrium (LD); however, here, we assume data have been pruned to approximately independent variants. Given $Z = \{Z_i\}_{i=1}^{n}$ and model parameters L, F, μ, τ, we can compute the likelihood as

$$\mathcal{L}(L, F, \mu, \tau | Z) = \prod\nolimits_{i=1}^{n} N\left(Z_i \middle| \sqrt{N_i}(Lf_i + \mu), \tau^{-1} I_p\right).$$

Consistent with probabilistic principal-component analysis (PCA) and similar approaches, we assume a standard normal prior over latent factors for each trait as $F \sim \prod_{i=1}^{n} N(f_i \mid 0, I_k)$. Next, to model our uncertainty in L, μ, we take a full Bayesian approach similar to a Bayesian PCA model.[16] Namely, we assume loadings for each SNP are sampled from a normal prior, $L \mid \alpha \sim \prod_{j=1}^{p} N(l_j \mid 0, \mathrm{diag}(\alpha)^{-1})$, where α is a $k \times 1$ vector reflecting the prior precision along each factor dimension. Similarly, we place a normal prior on the shared intercept $\mu \sim N(0, \varphi^{-1} I_p)$, where φ is the prior precision.

By modeling the intercept μ and loadings L as being sampled from normal distributions with precision parameters (φ and α, respectively), FactorGo shrinks estimates toward 0. Rather than require users to specify α *a priori*, we use ARD[16] to “shut off” uninformative factors, thus minimizing overfitting when k is misspecified, which is equivalent to placing a prior over α as $\alpha \sim \prod_k G(a_k, b_k)$, where the expected shrinkage effects $E(\alpha_k)$ on loadings are inferred from data. Altogether, FactorGo reflects a model where each SNP contributes to each latent dimension (albeit adaptively shrunk toward zero), and each trait has a representation across each latent dimension (albeit learned from the shrunken loadings projected onto the observed data).

Lastly, we place a prior over the shared residual variance across GWASs as $\tau \sim G(a_\tau, b_\tau)$ to capture the average residual variance due to non-linear genetic effect or shared environment across GWASs. We impose broad priors by setting hyperparameters $\varphi = a_k = b_k = a_\tau = b_\tau = 10^{-5}$.

Variational inference

Given our FactorGo model and observed Z-score summary data, we would like to infer the posterior distribution of parameters L, F, μ, α, τ. Unfortunately, there is no closed form expression for learning the posterior exactly, and thus, we leverage variational inference to infer an approximate posterior distribution.[16,17] Let D be the observed Z score and respective GWAS sample sizes. In brief, the true posterior distribution $P(L, F, \mu, \alpha, \tau \mid D)$ is approximated by a factorized tractable distribution from the conjugate families

$$P(L, F, \mu, \alpha, \tau \mid D) \approx Q_L(L \mid D)\, Q_F(F \mid D)\, Q_\mu(\mu \mid D)\, Q_\alpha(\alpha \mid D)\, Q_\tau(\tau \mid D),$$

where $Q_{\cdot}(\cdot)$ reflects a surrogate approximating posterior for individual model parameters. The optimal functional forms for each Q and respective variational parameters are identified by maximizing the evidence lower bound on the marginal likelihood (i.e., ELBO). During inference, variational parameters are updated iteratively until convergence. The model outputs estimates of posterior means and variances of L, F, μ, α, τ.

To further improve the scalability of our approach, we apply a parameter expansion design that converges more rapidly.[18] Namely, after each iteration step, the latent space F is centered using a weighted mean, and L is orthogonalized to reduce coupling effects of latent parameters (see Note S1). We implemented FactorGo in Python using just-in-time (JIT) compilation through the *JAX* package (see web resources), which generates and compiles heavily optimized C++ code in real time and operates seamlessly on CPU, GPU, or TPU (see data and code availability).

Simulations

To evaluate the performance of FactorGo and tSVD, we performed simulations under a polygenic additive model. Specifically, for i^{th} study, we generated a p-vector of true SNP effects β_i as a linear combination of k latent factors $\beta_i = Lf_i$, where the values of L, f_i were generated from $L_{jk} \sim N\left(0, \frac{h_g^2}{p \cdot k \cdot 2s_j(1-s_j)}\right)$ and $f_{ik} \sim N(0,1)$, where $j \in [p]$. The minor-allele frequency was sampled from $s_j \sim U(0.01, 0.5)$. For simplicity, we fixed the intercept μ to zeros. Given SNP heritability h_g^2, the total simulated variance in outcome Y was $Var(Y_i) = 1/h_g^2 * \sum_j (\beta_{ij} * 2s_j(1 - s_j))^2$. Then, residuals of each SNP effect in each study became $\sigma_{ij}^2 = Var(Y_i) - (\beta_{ij} * 2s_j(1 - s_j))^2$. Assuming the genotype was centered but not standardized, then the standard errors were $\widehat{SE_{ij}^2} = \sigma_{ij}^2 / \{N_i * (2s_j(1-s_j) + (2s_j)^2)\}$ on the per-allele unit, where GWAS sample size N_i was sampled empirically from 2,483 Pan-UKB studies in real data analysis (Figure S1). Finally, we added Gaussian noise to generate observed SNP effects $\widehat{\beta}_i \sim MVN_p(Lf_i, \widehat{\Sigma_i})$ for $i \in [n]$, where the diagonal values of $\widehat{\Sigma_i}$ were $\widehat{SE_{ij}}^2$. Observed Z-score summary statistics were calculated as $\widehat{\beta}_{ij} / \widehat{SE_{ij}}$.

For each simulated dataset, we applied tSVD and FactorGo on standardized observed Z-score matrices with size $n \times p$ to compare their reconstruction error on true latent parameters. Standardization was applied to columns such that each SNP vector had zero mean and unit variance. Assuming the true model was consistent with FactorGo model and the true number of latent factors k was known, we explored extensive scenarios by varying four different parameters: (1) number of traits n; (2) number of independent causal SNPs p; (3) number of true latent factors k; and (4) additive SNP heritability h_g^2 . Each simulated scenario has 30 replications. Next, we examined the influence of model misspecification under four conditions: (1) misspecified number of latent factors; (2) correlated standard errors due to GWAS sample overlap; (3) no latent factors (i.e., no pleiotropy) and only correlated standard errors; and (4) correlated test statistics due to moderate LD after LD pruning. Lastly, we examined the robustness of FactorGo across a grid of five hyperparameters regarding prior distributions.

Metrics for simulation

We evaluated the accuracy of FactorGo and tSVD across several metrics. First, to evaluate the accuracy in reconstructed SNP effects matrices $B = LF$, we calculated the Frobenius norm between estimates and ground truth, i.e., $\|B - \widehat{L}\,\widehat{F}\|_F$. For tSVD decomposition USV^T, we defined $\widehat{F} = US$ and $\widehat{L} = VS$. Second, we evaluated the accuracy in estimating variant loadings L and factor scores F. To account for rotation and scaling in inferred parameters, i.e., $(\widehat{L}R)(R^{-1}\,\widehat{F})$ can give the same data likelihood where $RR^{-1} = I$, we performed procrustes analysis to align the parameters with their ground truth. Briefly, given matrices A and B, procrustes analysis[19] aims to find a rotation matrix R and scaling s term such that $min_R\ \|A - sRB\Big|\|_F^2$ subject to $RR^{-1} = I$. Here, we applied procrustes analysis on the inferred loading matrix $\widehat{L}$ to learn an optimal rotation R and scaling factor s and then computed $\tilde{L} = \widehat{L}Rs$ and calculated a final reconstruction error as $\|L - \tilde{L}\|_F$. Using the same rotation matrix R and scaling factor s, we computed $\tilde{F} = (sR)^{-1}\widehat{F}$ and calculated reconstruction error as $\|F - \tilde{F}\|_F$.

When no latent factors existed and test statistics correlated across studies due to residual confounding, we applied Levene's test to compare the variance of inferred parameters. The motivation is that if non-zero error correlation induces false discovery of latent structures, then we expect the variance of $1/\ E(\alpha)$ (or eigenvalues) to deviate from the null of constant variance, i.e., $\sigma^2_{\rho^2=0} = \sigma^2_{\rho^2=0.1} = \ldots = \sigma^2_{\rho^2=1}$.

Quality control on traits from Pan-UKB

Out of the total 7,200 traits from up to 420,531 European individuals in the Pan-UKB (version 04/11/22; UK Biobank application number 68459; see web resources), we selected traits with number of cases >1,000 for binary traits and total sample size >1,000 for quantitative traits. Pan-UKB ran GWASs using scalable and accurate implementation of generalized mixed model (SAIGE) to obtain accurate p values for studies with a highly imbalanced ratio of case groups to control groups.[20] For continuous traits, we chose GWAS results under inverse rank normal transformation to correct for outcome distribution. For categorical traits, we selected disease outcomes (Table S1). As a result, the final list consisted of 1,677 binary and 806 quantitative traits (see manifest file in Table S6), spanning a wide spectrum of trait domains including diseases, medications, environmental exposures, physical and biomarkers measures, etc. We categorized all 2,483 traits into nine distinct groups based on the description of UKB field ID (Table S1). We observed marked differences in total sample size across traits, with mean 403,306 for binary traits and 183,577 for quantitative traits (Figure S1).

Quality control on genetic variants from Pan-UKB

We filtered ~28 million autosomal variants by INFO score >0.9 (imputation accuracy score), minor-allele frequency >1%, high quality (PASS variant in gnomAD), and high-confidence variants (not extremely rare variants) defined by Pan-UKB (Figure S2). Then, we excluded the human leukocyte antigens (HLAs) region (chr6:25,000,000–34,000,000 [hg19]), indels, and multi-allelic variants. To ensure pleiotropic components across variants, we included SNP variants associated with at least two traits using p-value threshold 5E−08. Lastly, we applied LD pruning through *Hail* software using the in-sample LD correlation matrix with window size of 250 kb and $r^2 < 0.3$ (see web resources). These quality check (QC) steps led to a Z-score data matrix of 51,399 variants by 2,483 traits. 0.002% missing values in Z scores were imputed using SNP means. For subsequent functional interpretation, we focused only on variants included in the 1,000 Genomes Project with functional annotation data[21] (see web resources).

Analyses of *Z*-score summary data

We implemented both FactorGo and tSVD to learn $k = 100$ latent factors and compare their findings. For FactorGo, we used broad priors by setting all hyperparameters to be 1E−05. For tSVD, we applied the *TruncatedSVD* function from *sklearn* python package with 20 iterations of randomized states (see web resources). The columns of Z-score data matrix in size $n \times p$ were centered and standardized. The inferred factors were ordered by variance explained in observed data for FactorGo (i.e., R^2) and by singular values for tSVD (see Note S1). To show robustness of inferred factors subject to the choice of k, we performed additional analysis using $k = 90$, 110, respectively, and compared the top two factors and three leading factors for focal traits in case studies.

Case studies

To validate results and discover biological insights, we highlighted four traits: BMI and standing height as characteristic polygenic traits, RA as a representative autoimmune disease (a family of diseases known to have substantial shared genetic basis), and PCa as the second most common cancer for men worldwide with under-explored shared architecture with other traits. For each trait, we characterized the three respective leading pleiotropic factors and compared results between FactorGo and tSVD.

Interpreting inferred parameters

To interpret the inferred parameters for latent factors and loadings, we transform estimates using previously described contribution and cosine scores.[7] To rank factors according to their relevance for a focal trait, we define the squared cosine score as

$$\cos^2_{k,i} = \tilde{F}^2_{k,i} \Bigg/ \sum_{k'} \tilde{F}^2_{k',i}$$

where $\tilde{F}_{k,i}$ is the posterior mean of the k th factor score for the i th trait, standardized by its posterior variance (to account for uncertainty around the mean estimate), i.e., $\tilde{F}_{k,i} = F_{k,i}/\sqrt{Var(F_{k,i})}$. This standardized contribution score upweights traits with greater sample size that provides more certainty (Figure S3). For this factor, we calculated contribution scores, respectively, defined as follows to rank all traits and all variants (Figure 1):

$$cntr_{k,i}{}^{phe} = \tilde{F}_{k,i}{}^2 \Bigg/ \sum_{i'} \tilde{F}_{k,i'}{}^2$$

$$cntr_{k,i}{}^{var} = L_{p,k}{}^2 \Bigg/ \sum_{p'} L_{p',k}{}^2$$

Higher contribution score means that the trait is better characterized by this factor or the variant has larger effect to this pleiotropic factor. To understand the shared biology characterized by a factor, we describe an approach to test for enrichment in functional annotations using factor loadings in the following section.

Enrichment analysis on variant loadings

To interpret shared biology characterized by inferred factors at the tissue- or cell-type resolution, we downloaded 205 LD score regression applied to specifically expressed genes (LDSC-SEG) annotations for variants in 1,000 Genomes Project[21,22] (see web resources). The annotations are genes specifically expressed in 205 tissue or cell types (e.g., brain vs. non-brain cell types). Because the variants were LD pruned to satisfy FactorGo model assumption, we leveraged LD score for these variants to collect tagging functional variants, which led us to use stratified LD-score regression (S-LDSC) software for annotation enrichment analysis. To leverage the machinery of S-LDSC[2,23] (see web resources) for identifying enriched annotation in variant factor loadings, we first transformed the loadings to Z-score scale. To achieve this, we defined a pseudo sample size for each factor as a weighted sum of GWAS sample sizes $N_k{}^{pseudo} = \sum_i cos_{k,i}{}^{2\ phe} \cdot N_i$. Then, we created a pseudo Z score by multiplying $\sqrt{N_k^{pseudo}} \cdot L_{j,k}$ as the Z-score input for S-LDSC software. This pseudo sample size $N_k{}^{pseudo}$ also specifies the sample size for LDSC. The LD scores were calculated using n = 489 European ancestry individuals from 1,000 Genomes with window size of 1 cM. Additionally, the LD scores for regression SNPs were calculated separately as the weight for S-LDSC.

We ran S-LDSC on loading-based Z scores against each annotation to identify enriched tissue or cell type (Figure 1), conditioning on baseline annotations described elsewhere.[22] We used flag *–n-blocks 4000* to obtain a more accurate standard error with 4,000 jackknife blocks instead of the default 200 because analyzed SNPs were LD pruned. We calculated q value to control factor-wise false discovery rate (FDR) <0.05 using the *qvalue* R package by fixing $\lambda = 0$, which is equivalent as Benjamini Hochberg adjusted p value (web resources). Note that the null distribution of p values from S-LDSC is not uniform because it is a one-sided test for positive coefficient, and thus it is not appropriate to estimate the proportion of null hypothesis using the q-value method.[24] To demonstrate that our S-LDSC approach is well calibrated, we created 10 non-overlapping annotations for randomly selected gene sets from ~20,000 genes and computed the enrichment of these annotations over all factors at FDR <5%. To compute the specificity of enriched tissue or cell types between inferred factors, we calculated all pairwise Jaccard indexes. Briefly, the Jaccard index measures the similarity between two sets A, B by $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, which is the ratio of the number of shared elements over the total number of unique elements.

LDSC analysis for leading traits

To illustrate the benefit of learning shared genetic components using FactorGo compared with pairwise analysis of traits, we first examined how factor scores between trait pairs reflect their genetic correlation. For each of the 20 leading traits linked to a focal trait in its leading factor, we calculated their factor score correlation and genetic correlation using LDSC. To showcase the consistency or difference in enrichment analysis between joint model and single-trait analysis, we repeated the LDSC enrichment analysis using genome-wide variants for each of the 20 leading traits for each leading factor.

Results

Method evaluation in simulations under model assumptions

We assessed the performance of FactorGo in learning latent parameters across different simulated genetic architectures and compared results with tSVD as a baseline.

First, we found that FactorGo outperformed tSVD, exhibiting lower reconstruction error in trait factor scores F across all simulated scenarios (Wilcoxon p = 3.64E−109; Figures 2A and S4). Moreover, we observed the FactorGo error in trait factor scores F decreased with the increasing number of traits (p = 2.09E−24; Figure S4A) and number of true latent factors (p = 7.30E−26; Figure S4C). Error in F remained roughly constant across varying numbers of causal SNPs (p = 0.99; Figure S4B) and average SNP heritability (p = 0.36; Figure S4D).

Second, although error of variant loading L was not significantly different between FactorGo and tSVD (p = 0.29; Figure 2B), we found FactorGo error decreased with increasing number of traits (p = 5.22E−15; Figure S4A), number of true latent factors (p = 1.40E−23; Figure S4C), and average SNP heritability (p = 0.071; Figure S4D). The error in loadings increased with increasing causal SNPs (p = 8.40E−06; Figure S4B). The accuracy in genetic effect B estimation was not statistically different between FactorGo and tSVD (p = 0.10; Figure 2C).

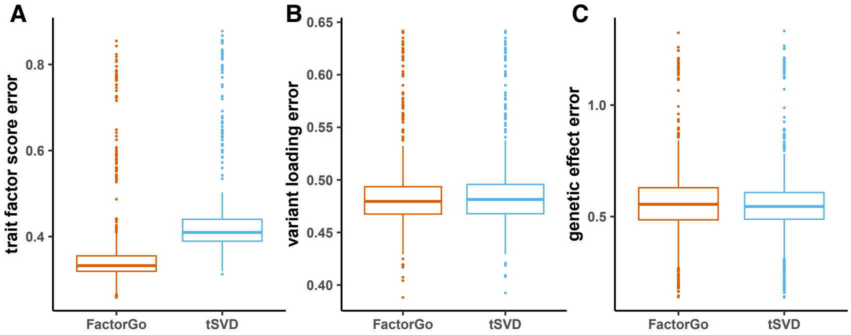


Figure 2. FactorGo provides accurate estimates of model parameters

(A–C) We report errors for (A) trait factor score F, (B) variant loading L, and (C) genetic effect B aggregated over four sets of simulations letting varying either the number of studies (n), the number of SNPs (p), the number of true latent factors (k), and SNP heritability (h_g^2) (see separate results in Figure S4). The median value is displayed as a band inside each box. Boxes denote values in the second and third quartiles. The length of each whisker is 1.5 times the interquartile range. All values lying outside the whiskers are considered to be outliers.

Overall, our simulations demonstrate FactorGo provides similar estimates of model parameters as tSVD, with a significant improvement of trait factor scores.

Method evaluation in simulations under model misspecification

Next, we sought to assess the performance of FactorGo under various settings reflecting model misspecification. First, we investigated when the specified k differs from the true number of latent factors. When the true number of latent factors $k = 10$, FactorGo performed similarly as tSVD in estimating trait factor scores F across varying k from 2 to 20 (p = 0.21; Figure 3A). However, FactorGo provided more accurate estimates in trait factor scores F than tSVD (p = 0.027) when k is underspecified ($k < 10$) compared with when k is overspecified ($k > 10$; Figure 3A). For variant loading L, the error was not significantly different between FactorGo and tSVD (p = 0.25; Figure 3B). Interestingly, the estimates for genetic effects B were more accurate in FactorGo (p = 0.047) across different k, especially when k was overestimated (p = 2.48E−17; Figure 3C).

Second, when standard errors and test statistics are correlated due to non-zero LD between SNPs, we observed that FactorGo consistently outperformed tSVD in reconstructing trait factor scores (p = 3.04E−78; Figures S5A–S5C). FactorGo was robust across varying magnitudes of correlated standard errors in estimating trait factor scores (p = 1.00) and variant loadings (p = 0.93; Figure S5A), whereas their combined predicted effects were less resilient. Importantly, when ρ_e^2 matched values estimated from real data (average 0.057; SD = 0.25; Figure S6A), we observed a less-pronounced effect on inferential bias (Figure S6B). Third, when no latent factors exist and correlated standard errors across traits due to unmeasured confounding (i.e., shared environment), we found little evidence of latent factor signals in $1/E(\alpha)$ from FactorGo (p = 1.00) or eigenvalues from tSVD (p = 1.00; Figure S5D), suggesting both approaches are robust to this confounding.

Lastly, we evaluated the sensitivity of FactorGo to choices of five hyperparameters involved with α (i.e., prior loading variance), μ (i.e., average SNP effect), and τ (i.e., residual heterogeneity). For each of the scenarios, we found FactorGo was robust to varying choices of these values in estimating true effects (p = 0.96), trait factor scores (p = 0.93), and variant loadings (p = 0.90; Figure S7).

Overall, our simulation results demonstrate that FactorGo accurately identifies latent representation of traits when k is underestimated, when test statistics across SNPs are correlated due to LD, and when standard errors are correlated across traits due to unmeasured confounding (i.e., shared environment).

FactorGo improves interpretation of the pleiotropic components of 2,483 UKB traits

Having demonstrated the performance of FactorGo in simulations, we next characterized 100 pleiotropic factors of 2,483 real traits from the Pan-UKB (mean N = 331,980; see web resources). We selected traits by their case group or total sample size >1,000. Initial screening on ~28 million variants by INFO >0.9 and minor-allele frequency >1% resulted in 8,449,689 high-quality common variants. We retained 7,624,608 biallelic non-HLA SNP variants and found 1,037,929 of them associated with at least two traits at p value < 5E−08. Next, we subsetted to 1,023,655 variants with LDSC-SEG annotation data followed by LD pruning with window size of 250 kb and $r^2 < 0.3$. Finally, we constructed a matrix of GWAS Z scores at 51,399 non-HLA LD-pruned SNP variants across each of the 2,483 traits (see material and methods). On average, each GWAS trait has 109 (SD = 541) significant variants. We applied FactorGo and tSVD to the QCd Z-score matrices to learn 100 pleiotropic factors. Both methods required approximately the same amount of runtime (~10 min for FactorGo on 2 GPUs; Figure S8) and explained similar amounts of variance in observed data (38.07% vs. 37.76%). For each method, we ranked factors by the proportion of variance explained. For FactorGo, we confirmed the robustness of posterior variance estimates by observing the entropy of posterior covariance was smaller for traits with larger sample size (Figure S9).

First, we reported the projection of all traits over the top two FactorGo pleiotropic factors in Figure 4. Factor 1 was driven by body weight and basal metabolic rate, and factor 2 was driven by human standing height. We obtained similar patterns for tSVD factors (Figure S10). Interestingly,

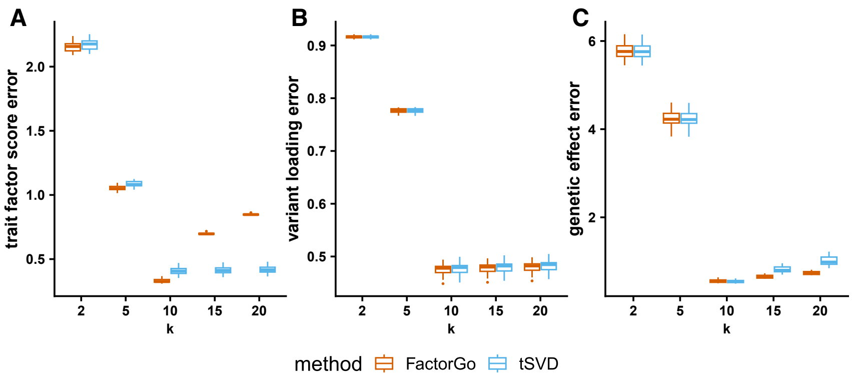


Figure 3. FactorGo outperforms tSVD in trait factor scores when k is underspecified
(A–C) We report reconstruction error for (A) trait factor score F, (B) variant loading L, and (C) genetic effect B in simulations under varying user-defined latent dimensions $k = 2, 5, 10, 15, 20$ when fixing true $k = 10$ (and $p = 2000$, $n = 100$, and $h_g^2 = 0.1$).

only FactorGo implied the shared comorbidity of COVID-19 with BMI-related traits in factor 1, an association that has been reported previously.[25] Characterization of factors 1 and 2 is given in the section “characterizing shared biology in FactorGo pleiotropic factors” below. Other leading factors were primarily driven by traits with higher heritability compared with factors that explained less Z-score variance (p = 1.99E−17 and 2.99E−18, respectively; Figure S11), which is consistent with heritability reflecting variation in allelic effect sizes. Additionally, as a proof of concept, we showed that the factor score correlation between leading trait-focal trait pairs tracked closely with their genetic correlation for four focal traits discussed later (Figure S12), which validated that FactorGo model effectively decomposed the genetic correlation across traits. This consistency decayed for factors in lower rank (F55, F86) as there was less variation explained by those factors.

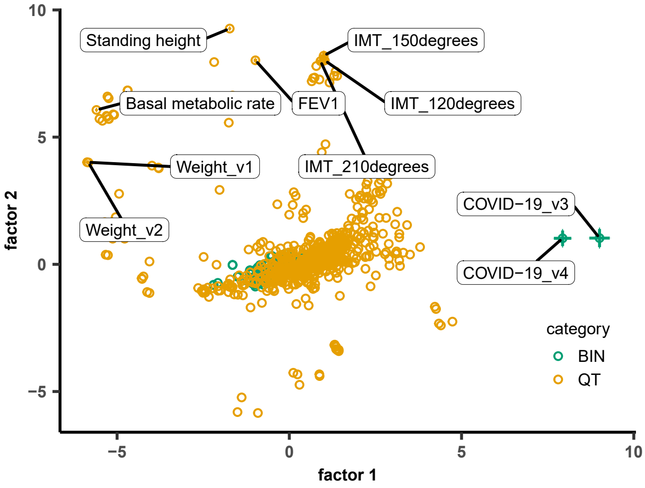


Figure 4. The top two factors in FactorGo characterize traits involved with body weight and height, respectively
We report the projection of 2,483 UK Biobank traits over the top two FactorGo pleiotropic factors. Error bars were 2 times the square root of posterior variance for trait factor scores and plotted only for highlighted traits. Binary (BIN) and quantitative (QT) traits were colored differently. FEV1, forced expiratory volume in 1 second; IMT, mean carotid intima-medial thickness; Weight_v1, amalgamated measure of weight by multiple means; Weight_v2, weight measured during impedance measurement. COVID-19_v3 and v4, tested for COVID-19 positive in two different waves.

Second, by quantifying and ranking the relative importance of pleiotropic factors related to a trait using squared cosine scores (see material and methods), we observed that the cumulative squared cosine score for each trait was higher in FactorGo than in tSVD at each rank of pleiotropic factor ($p < 0.05/99$; Figure S13). To evaluate the sufficiency of these 100 factors in explaining genetic associations from observed data, we found the variance explained by each factor leveled off quickly for both FactorGo and tSVD (Figure S14A). The posterior mean of prior precision parameter α tracked closely with the variance explained by each factor, suggesting that FactorGo successfully shrunk less-informative factors (Figure S14B). Finally, to show robustness of FactorGo results with respect to choice of k, we performed additional analysis using $k = 90$ and 110. The top two latent factors were highly consistent in 20 leading traits and 10 leading variants across $k = 90$, 100, and 110 results (Figure S15).

Third, we evaluated the ability of FactorGo and tSVD to identify relevant shared biology demonstrated by computing tissue-specific enrichment of factor-specific loadings using S-LDSC (see material and methods; we note that this method was well calibrated under FDR <5%; Figure S16). Overall, we found that the S-LDSC coefficient Z statistics were higher in FactorGo compared with those from tSVD (mean 0.051 vs. −0.042, p = 2.58E−10; Figure S17). Of the 100 FactorGo factors, we observed that 69 were enriched with at least one tissue or cell type at factor-wise FDR <5%, in contrast with only 40 when using tSVD. FactorGo factors were enriched with seven tissue or cell types, on average, and spanned 191/205 tissue or cell types compared with 130/205 from tSVD (p = 6.59E−13). To show specificity of enriched tissue or cell types between inferred factors, we calculated all pairwise Jaccard indexes and found the mean similarity for FactorGo is 0.030, which is lower than 0.045 in tSVD (p = 9.37E−04).

Altogether, our results demonstrate that FactorGo identifies biologically meaningful pleiotropic components at the tissue- and cell-type resolution.

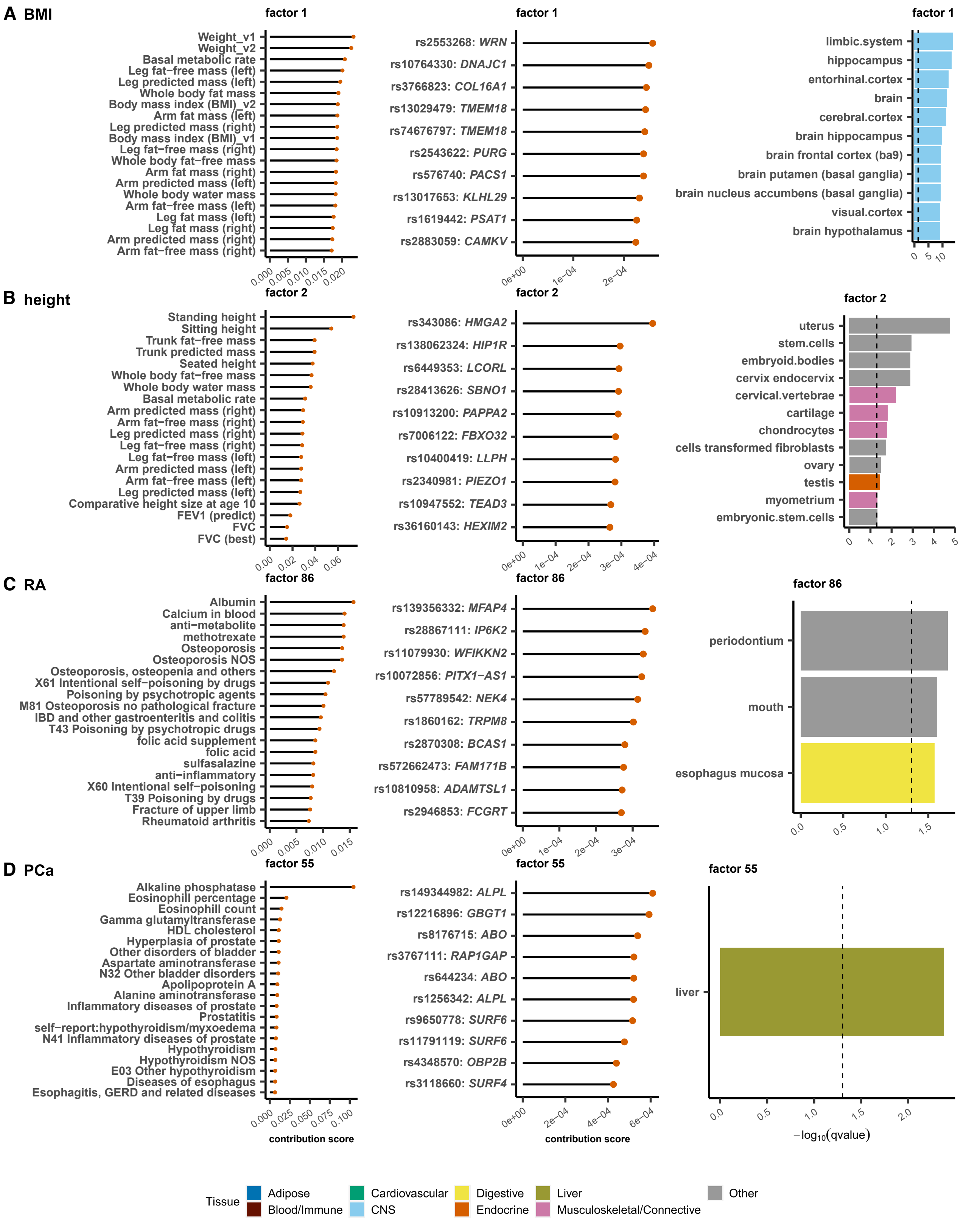


Figure 5. Characterizing shared biology in pleiotropic factors leading four representative traits

(A–D) We characterized the pleiotropic factors leading (A) BMI, (B) height, (C) rheumatoid arthritis (RA), and (D) prostate cancer (PCa). For each focal trait (row), we identified its leading factor and reported the contribution scores of the 20 leading traits of this factor, the 10 leading variants with their closest gene, and $-log_{10}(qvalue)$ at FDR <5% (symbolized by dashed vertical line) for significantly enriched

(legend continued on next page)

Characterizing shared biology in FactorGo pleiotropic factors

To characterize the pleiotropic factors identified by FactorGo, we analyzed the leading factors of four representative traits: BMI, height, RA, and PCa. For each trait, we identified its most relevant factor using squared cosine scores, identified the other traits leading this factor using contribution scores, identified the genetic variants leading this factor using contribution scores, and characterized the biology of this factor using S-LDSC on 205 tissue- and cell-type-specific annotations (see material and methods). To ensure that inferred loadings meet the assumptions required for valid S-LDSC analyses, we inspected the linear relationship between LD scores and transformed loadings (pseudo Z scores) for four leading factors associated with four focal traits and found consistent representation (Figure S18), thus supporting the validity of our approach. We assessed that our results were overall consistent across $k = 90$, 100, 110 (Figures S19–S22).

BMI is characterized by factor 1, associated to brain cell types

The leading factor for BMI was factor 1 (squared cosine score: 58.85%), which was characterized by body weight (contribution score: 2.32%), basal metabolic rate (2.08%), and body fat masses (cumulative 17.74% across 13 traits; Figure 5A; Table S2). The leading variants were proximal to genes such as *WRN* associated with Werner Syndrome (and thus short stature and abnormal fat distribution[26]; rs2553268:G>T: 0.026%) and *TMEM18* associated with obesity (rs13029479:G>A: 0.024%; rs74676797:G>A: 0.024%).[27] Out of the 33 tissues and cell types significantly enriched in factor 1, 31 were brain cell types including the limbic system and hippocampus (Figure 5A), which is consistent with previous findings of brain-specific enrichments in BMI genetic data.[2,22] This brain-specific enrichment was also concordant with leading trait enrichment analysis (Figure S23). The next two leading factors for BMI (factors 4 and 7) identified its shared biology with pharynx and digestive tissues, respectively (Note S2; Figure S24). We performed the same analysis using results from tSVD and found no enrichment of cell types in the leading factor for BMI, despite similarly characterized body fat traits (Figure S25).

Standing height is characterized by factor 2, associated with musculoskeletal tissues

As the leading factor for standing height, factor 2 (squared cosine score: 38.67%) characterized leading traits as standing height (7.36%), sitting height (5.41%), and body fat masses (1.39%; Table S2). These associations were driven primarily by an intron variant in height-associated gene *HMGA2* (rs343086:T>C: 0.04%).[28,29] As expected, factor 2 exhibited enrichment for musculoskeletal tissues such as cartilage and chondrocytes (Figure 5B). Additionally, we replicated enrichment for reproductive organs such as uterus and cervix.[22,30] This result is also consistent with prior work demonstrating that overexpression of *HMGA2* alters production of growth hormone in mice[31] in addition to reproductive tissue development.[32] These enrichments in factor 2 were also concordant with leading trait enrichment analysis (Figure S26). The next two leading factors for height suggested a shared biology with cardiovascular tissues and immunity, respectively (Note S2; Figure S27). For tSVD, we found its leading factor similarly characterized height traits but did not exhibit evidence of cell-type enrichment (Figure S28).

RA leading factor is driven by inflammatory mechanisms

For RA, factor 86 (squared cosine score: 7.17%) was explained primarily by inflammation-related traits (Figure 5C) such as blood albumin level (1.57%), blood calcium level (1.40%), methotrexate (a common treatment for RA; 1.39%), osteoporosis conditions (cumulative 5.52% across five traits; Table S3), and other autoimmune diseases such as inflammatory bowel disease (0.96%).[33–35] We found these signals were driven by variants proximal to genes *MFAP4* (rs139356332:G>C: 0.036%) and *IP6K2* (rs28867111:G>A: 0.033%), both of which are involved with inflammatory mechanisms.[36,37] Interestingly, we observed factor 86 exhibited enrichment in periodontium and mouth (Figure 5C), which is supported by prior epidemiological evidence of common periodontal conditions in individuals with RA due to autoantibodies and arthritis triggered by oral pathogens.[38] Interestingly, these enrichments in factor 86 were not found in single-trait enrichment analysis (Figure S29), which is likely caused by underpowered disease traits in biobank studies. Our selection in variants includes the pleiotropic variants with the strongest signals that can be overwhelmed by the genome-wide underpowered background in single-trait analysis. The next two leading factors for RA (factors 75 and 76) suggested a shared biology with mechanisms in the kidney, liver, and central nervous system (Note S2; Figure S30). Different from FactorGo, the leading factor for RA from tSVD characterized insulin-like growth factor 1 (IGF-1) measure and cardiac disorders but not enriched with any cell types (Figure S31).

PCa leading factor identifies ALP as a PCa candidate biomarker

For PCa, the leading factor was factor 55 (squared cosine score: 17.94%), characterized by diseases in prostates,

LDSC-SEG annotations (truncated to 10 if more than 20 enriched annotations). See detailed result in Table S4. FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; Weight_v1, amalgamated measure of weight by multiple means; Weight_v2, weight measured during impedance measurement; BMI_v1, BMI estimated by impedance measurement; BMI_v2, BMI estimated based on weight and height; NOS, not otherwise specified.

including hyperplasia of prostates (1.13%) and inflammatory diseases in prostates (1.63%) (Figure 5D). The leading trait was ALP level in blood (10.47%), associated to the leading missense variant in *ALPL* (c.224G>A [GenBank: NM_001177520.3] [p.Arg75His] [rs149344982: 0.061%]). Because ALP is an enzyme mostly produced by the liver and bone, this factor was indeed enriched with genes specifically expressed in the liver. Previous work found higher serum ALP was associated with poor overall survival rate of individuals with PCa, which likely reflects bone metastatic tumor load.[39] Similarly, liver enrichment in factor 55 was consistent with leading trait enrichment analysis (Figure S32). The next two leading factors for PCa (factors 1 and 58) suggested shared comorbidities of PCa involved with BMI and hormonal disorders (Note S2; Figure S33), which is consistent with previous works investigating dietary risk factors[40] as well as the well-documented role of hormonal dependency due to expression of androgen receptor.[41] Different from FactorGo, the leading factor for PCa from tSVD prioritized corneal resistance factors, geographic home locations, and heel bone measures (Figure S34). Additionally, tSVD results displayed enrichment for genes expressed specifically in colon, suggesting alternative shared biological mechanisms compared with FactorGo.

Discussion

In this work, we presented FactorGo to identify and characterize pleiotropic components across thousands of human complex traits and diseases using *Z*-score summary statistics. Our method enables investigating the phenome-wide shared genetic components while appropriately modeling uncertainty in variant effect estimates. When applied to 2,483 phenotypes from the UKB individuals, we found that FactorGo factors explained more variance on average and were more powerful in identifying shared biology compared with tSVD factors. We validated brain-specific enrichment for BMI factors as well as muscular-skeletal and reproduction enrichment for height factors. For disease traits, FactorGo suggests a shared etiology between RA and periodontal conditions. Moreover, we found ALP as a candidate but less-established biomarker for PCa, which provided evidence for further experimental validation.

FactorGo has several advantages compared with the scalable but model-free approach tSVD. First, FactorGo learns pleiotropic factors at similar computational cost by leveraging state-of-the-art variational inference and fast python implementation. Second, we showed using simulations that FactorGo outperformed tSVD in estimating trait factor score under model assumption and model misspecification such as correlated standard errors due to GWAS sample size. Third, in real data analyses, we found more enrichment of tissue or cell types in FactorGo factors than in tSVD factors.

We note that the aims of FactorGo are similar with recent approaches seeking to partition shared and distinct genetic architectures across multiple traits using GWAS summary data. First, tSVD applies a matrix decomposition on the observed *Z*-score summary statistics matrix directly to identify latent genetic components,[7] whereas FactorGo seeks decomposition of the true genetic effects by modeling the uncertainty around genetic effect estimates. Second, GenomicSEM is a flexible framework that identifies SNPs with effects specific to one or a subset of traits.[12] In contrast, the SNP effect on traits in the FactorGo model is only through factors that characterize shared genetic liability across relevant traits, where the relevancy is ranked by trait factor scores. Lastly, pleiotropic decomposition regression (PDR) parses apart different underlying SNPs across factors that characterize putative mechanisms by a parsimonious decomposition.[10] In contrast, FactorGo attempts a parsimonious explanation by employing an ARD prior that penalizes factor loadings plus orthogonalization of factors under parameter expansion design. Overall, FactorGo provides a scalable probabilistic framework to characterize the latent genetic components shared across human complex traits by leveraging widespread pleiotropy.

Our tool has several implications for downstream analyses. First, we demonstrated that analyzing phenome-wide GWAS summary statistics from biobanks can not only recapitulate known shared biology for traits such as BMI, height, and RA but also nominate candidate biomarkers in diseases for further clinical evaluation such as ALP for PCa. This testifies the benefit of enabling scalability of model-based statistical approaches jointly analyzing thousands of GWAS summary data from large biobanks. Second, leveraging factor loadings within enrichment analysis using differentially expressed gene annotations allowed us to interpret the biology of a given factor at tissue- or cell-type level. Our application of S-LDSC to variant loadings readily allows analyzing other functional annotations such as chromatin accessibility and transcriptional factors. In theory, FactorGo can be applied to phenotype matrix, which leads to a decomposition of phenotypic, rather than genetic, correlations. Here, the latent factors underlying phenotypic correlation reflect both shared environment and genetics, which can be used as input for downstream Factor GWAS analysis. However, we note that working with thousands of phenotypes from hundreds of thousands of individuals requires greater computational overhead.

Although FactorGo has provided robustness in simulations and rich insights in the analyses of UKB phenotypes, it has some limitations. First, our method focused on learning pleiotropic factors from linear genetic effects and ignored non-linear or epistatic effects. While many lines of evidence pointed to linear models capturing the bulk of trait heritability,[42,43] our results also illustrated rich meaningful biological insight that could be obtained from linear effects alone. Second, our model assumes independence of residual errors, which was unlikely to be true given overlapped samples in large biobank GWASs. However, we showed in simulation that the estimation of latent parameters was robust to error correlation. Third, FactorGo didn't outcompete tSVD in estimating variant loadings in our simulations. However, we provided a probabilistic model to account for heterogeneity in summary statistics across GWASs without adding extra run-time cost. Fourth, while our method requires predefining the number of latent factors k, our simulations have shown that results are biased if k was fixed to a too-high value. However, to ensure that this limitation is unlikely to impact our results, we performed additional analysis using $k = 90$ and 110. The top two latent factors were highly consistent in 20 leading traits and 10 leading variants across $k = 90$, 100, and 110 results (Figure S15). The leading factors for BMI, height, RA, and PCa were overall consistent in traits (Figures S19–S22). Fifth, in real data analysis, our selection of variants using genome-wide significance thresholds can underestimate the degree of pleiotropy due to lack of power, especially in disease traits. For example, in the case study of PCa, we did not observe PCa in the top rank of leading factors, suggesting either PCa has limited shared components with other traits or lack of power in GWASs to estimate the variant effects. Despite this, we were still able to recapitulate known shared biology for BMI, height, and RA using this subset of pleiotropic variants. Similarly, our selection of variants involved an LD-pruning procedure. While pruning could limit the functional interpretation of the latent factors, our gene-set analyses leveraging LD scores computed on a sequenced reference panel mitigate this issue. We anticipate that improvement in fine-mapping techniques and ongoing efforts to perform fine mapping on hundreds of phenotypes at the biobank scale[44] should improve variant selection in the near future. Sixth, FactorGo factors are identifiable only up to sign, which makes interpretation challenging (e.g., risk increasing/decreasing). Here, we validated biological interpretability of factors using enrichment analysis for traits with better-understood genetic components such as height and BMI. Despite this limitation, FactorGo factors estimated from phenome-wide data can help generate hypotheses for experimental validation. Seventh, unlike other methods based on non-negative matrix factorization,[4] our model did not distinguish between varying directional effects of pleiotropic factors but rather focused on non-directional summary of pleiotropic effects. Eighth, recent works have highlighted that shared effect sizes across traits might be driven by assortative mating.[45] Further investigation is required to see how it impacts the interpretation of our results. Lastly, although our method was developed for single-ancestry analysis, it can be extended to multi-ancestry data and learn shared genetic components. Taking it a step further regarding the model and subsequent interpretation, it is also possible to incorporate functional annotation as priors so that interpreting functional enrichment *a posteriori* is more straightforward.

In conclusion, FactorGo provides a variational Bayesian factor analysis model on GWAS summary statistics to learn and characterize pleiotropic factors across thousands of human complex traits and diseases. It allows rich biological interpretation at tissue- or cell-type-specific level.

Data and code availability

- The accession number to GWAS summary statistics and results reported in this paper are available on Zenodo: https://zenodo.org/record/7765048
- Original GWAS summary statistics are available on AWS cloud. Please see details on website of Pan-UKB (https://pan.ukbb.broadinstitute.org/) and download links in Table S6.
- In-sample LD correlation matrix for Europeans released by Pan-UKB is available on AWS cloud: s3a://pan-ukb-us-east-1/ld_release/UKBB.EUR.ldadj.bm
- FactorGo software: https://github.com/mancusolab/FactorGo
- FactorGo analysis code: https://github.com/mancusolab/FactorGo_analysis

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.09.015.

Acknowledgments

This work was funded in part by the National Institutes of Health (NIH) under awards P01CA196569, R01HG012133, R01CA258808, R35GM147789, and R00HG010160.

We thank Dr. Tiffany Amariuta for her comments and suggestions to this manuscript.

Author contributions

Z.Z., S.G., and N.M. developed the method. Z.Z. performed analysis. J.J., N.S., and A.K. prepared data and performed analyses. All authors edited and approved of the manuscript.

Declaration of interests

N.M. is a member of the HGG Advances Editorial Board.

Received: March 27, 2023
Accepted: September 27, 2023
Published: October 24, 2023

Web resources

Hail: https://hail.is/docs/0.2/
JAX: https://github.com/google/jax
LDSC-SEG annotations: https://alkesgroup.broadinstitute.org/LDSCORE/LDSC_SEG_ldscores/
Pan-UK BioBank: https://pan.ukbb.broadinstitute.org/
qvalue R package: https://github.com/StoreyLab/qvalue
TruncatedSVD: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html
1,000 Genome annotations: https://alkesgroup.broadinstitute.org/LDSCORE/

References

1. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348.
2. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
3. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
4. Udler, M.S., Kim, J., von Grotthuss, M., Bonàs-Guarch, S., Cole, J.B., Chiou, J., Christopher, D., Anderson on behalf of METASTROKE and the ISGC, Boehnke, M., Laakso, M., Atzmon, G., et al. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. *15*, e1002654.
5. Wang, L., Balmat, T.J., Antonia, A.L., Constantine, F.J., Henao, R., Burke, T.W., Ingham, A., McClain, M.T., Tsalik, E.L., Ko, E.R., et al. (2021). An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. Genome Med. *13*, 83.
6. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.
7. Tanigawa, Y., Li, J., Justesen, J.M., Horn, H., Aguirre, M., DeBoever, C., Chang, C., Narasimhan, B., Lage, K., Hastie, T., et al. (2019). Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. Nat. Commun. *10*, 4064.
8. Chasman, D.I., Giulianini, F., Demler, O.V., and Udler, M.S. (2020). Pleiotropy-Based Decomposition of Genetic Risk Scores: Association and Interaction Analysis for Type 2 Diabetes and CAD. Am. J. Hum. Genet. *106*, 646–658.
9. He, Y., Chhetri, S.B., Arvanitis, M., Srinivasan, K., Aguet, F., Ardlie, K.G., Barbeira, A.N., Bonazzola, R., Im, H.K., et al.; GTEx Consortium (2020). sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression. Genome Biol. *21*, 235.
10. Ballard, J.L., and O'Connor, L.J. (2022). Shared components of heritability across genetically correlated traits. Am. J. Hum. Genet. *109*, 989–1006.
11. Dahl, A., Iotchkova, V., Baud, A., Johansson, Å., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A., and Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. Nat. Genet. *48*, 466–472.
12. Grotzinger, A.D., Rhemtulla, M., de Vlaming, R., Ritchie, S.J., Mallard, T.T., Hill, W.D., Ip, H.F., Marioni, R.E., McIntosh, A.M., Deary, I.J., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat. Hum. Behav. *3*, 513–525.
13. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
14. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27*, S2–S8.
15. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518.
16. Bishop, C.M. (1999). Variational principal components, pp. 509–514.
17. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. *112*, 859–877.
18. Luttinen, J., and Ilin, A. (2010). Transformations in variational Bayesian factor analysis to speed up learning. Neurocomputing *73*, 1093–1102.
19. Gower, J.C. (1975). Generalized procrustes analysis. Psychometrika *40*, 33–51.
20. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
21. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
22. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.
23. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
24. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445.
25. Gao, M., Piernas, C., Astbury, N.M., Hippisley-Cox, J., O'Rahilly, S., Aveyard, P., and Jebb, S.A. (2021). Associations between body-mass index and COVID-19 severity in 6·9 million people in England: a prospective, community-based, cohort study. The Lancet Diabetes & Endocrinology *9*, 350–359.
26. Muftuoglu, M., Oshima, J., von Kobbe, C., Cheng, W.-H., Leistritz, D.F., and Bohr, V.A. (2008). The clinical characteristics of Werner syndrome: molecular and biochemical diagnosis. Hum. Genet. *124*, 369–377.
27. Landgraf, K., Klöting, N., Gericke, M., Maixner, N., Guiu-Jurado, E., Scholz, M., Witte, A.V., Beyer, F., Schwartze, J.T., Lacher, M., et al. (2020). The Obesity-Susceptibility Gene TMEM18 Promotes Adipogenesis through Activation of PPARG. Cell Rep. *33*, 108295.
28. Yang, T.-L., Guo, Y., Zhang, L.-S., Tian, Q., Yan, H., Guo, Y.-F., and Deng, H.-W. (2010). HMGA2 is confirmed to be associated with human adult height. Ann. Hum. Genet. *74*, 11–16.
29. Weedon, M.N., Lettre, G., Freathy, R.M., Lindgren, C.M., Voight, B.F., Perry, J.R.B., Elliott, K.S., Hackett, R., Guiducci, C., Shields, B., et al. (2007). A common variant of HMGA2 is

associated with adult and childhood height in the general population. Nat. Genet. *39*, 1245–1250.

30. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J. 'an, Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
31. Fedele, M., Battista, S., Kenyon, L., Baldassarre, G., Fidanza, V., Klein-Szanto, A.J.P., Parlow, A.F., Visone, R., Pierantoni, G.M., Outwater, E., et al. (2002). Overexpression of the HMGA2 gene in transgenic mice leads to the onset of pituitary adenomas. Oncogene *21*, 3190–3198.
32. Lee, M.O., Li, J., Davis, B.W., Upadhyay, S., Al Muhisen, H.M., Suva, L.J., Clement, T.M., and Andersson, L. (2022). Hmga2 deficiency is associated with allometric growth retardation, infertility, and behavioral abnormalities in mice. G3 (Bethesda). *12*, jkab417.
33. Yang, B., Gross, M.D., Fedirko, V., McCullough, M.L., and Bostick, R.M. (2015). Effects of calcium supplementation on biomarkers of inflammation and oxidative stress in colorectal adenoma patients: a randomized controlled trial. Cancer Prev. Res. *8*, 1069–1075.
34. Don, B.R., and Kaysen, G. (2004). Serum albumin: relationship to inflammation and nutrition. Semin. Dial. *17*, 432–437.
35. Ginaldi, L., Mengoli, L.P., and De Martinis, M. (2009). Osteoporosis, Inflammation and Ageing. In Handbook on Immunosenescence: Basic Understanding and Clinical Applications, T. Fulop, C. Franceschi, K. Hirokawa, and G. Pawelec, eds. (Springer Netherlands)), pp. 1329–1352.
36. Kanaan, R., Medlej-Hashim, M., Jounblat, R., Pilecki, B., and Sorensen, G.L. (2022). Microfibrillar-associated protein 4 in health and disease. Matrix Biol. *111*, 1–25.
37. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. Nat. Genet. *44*, 502–510.
38. Konig, M.F., Abusleme, L., Reinholdt, J., Palmer, R.J., Teles, R.P., Sampson, K., Rosen, A., Nigrovic, P.A., Sokolove, J., Giles, J.T., et al. (2016). Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. Sci. Transl. Med. *8*, 369ra176.
39. Li, D., Lv, H., Hao, X., Hu, B., and Song, Y. (2018). Prognostic value of serum alkaline phosphatase in the survival of prostate cancer: evidence from a meta-analysis. Cancer Manag. Res. *10*, 3125–3139.
40. Salem, S., Salahi, M., Mohseni, M., Ahmadi, H., Mehrsai, A., Jahani, Y., and Pourmand, G. (2011). Major dietary factors and prostate cancer risk: a prospective multicenter case-control study. Nutr. Cancer *63*, 21–27.
41. Lindström, S., Finucane, H., Bulik-Sullivan, B., Schumacher, F.R., Amos, C.I., Hung, R.J., Rand, K., Gruber, S.B., Conti, D., Permuth, J.B., et al. (2017). Quantifying the Genetic Correlation between Multiple Cancer TypesThe Genetic Correlation between Multiple Cancer Types. Cancer Epidemiol. Biomarkers Prev. *26*, 1427–1435.
42. Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. *4*, e1000008.
43. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.
44. Kanai, M., Elzur, R., Zhou, W., Finucane, H.K., Global Biobank Meta-analysis Initiative, and Daly, M.J. (2022). Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell Genom. *2*, 100210.
45. Border, R., Athanasiadis, G., Buil, A., Schork, A.J., Cai, N., Young, A.I., Werge, T., Flint, J., Kendler, K.S., Sankararaman, S., et al. (2022). Cross-trait assortative mating is widespread and inflates genetic correlation estimates. Science *378*, 754–761.

The American Journal of Human Genetics, Volume *110*

Supplemental information

A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics

Zixuan Zhang, Junghyun Jung, Artem Kim, Noah Suboc, Steven Gazal, and Nicholas Mancuso

Supplemental Note 1

1 Overview

Conventional factor analysis model decomposes the variance-covariance matrix of observed data into common variance due to shared latent factors and specific variance[1]. Both common and specific variance are estimated from data. To leverage the uncertainty estimates SE^2 in effect sizes from GWAS summary statistics, we extended the conventional factor analysis to use observed standard error as specific variance in Gaussian distribution. Taking a step further under reasonable assumptions, we simplified this SNP effect model to Z-score model (FactorGo). Under a Bayesian hierarchical model, FactorGo leverages variational inference to infer posterior moments of latent parameters. We applied a parameter expansion design to substantially accelerate convergence rate.

2 FactorGo model

2.1 Notation

We denote matrices in uppercase bold (e.g., $\mathbf{X}$), vectors in lowercase bold (e.g., $\mathbf{x}$), and scalars in italicized lowercase (e.g., x). Further, we denote the i^{th} column of matrix $\mathbf{X}$ as column-vector $\mathbf{x}_i$, and the j^{th} row of matrix $\mathbf{X}$ as as column-vector $\mathbf{x}^j$. We denote the transpose of a matrix as $\mathbf{X}^\intercal$ and vector as $\mathbf{x}^\intercal$.

2.2 Model

Let $\hat{\boldsymbol{\beta}}_i$ be a $p \times 1$ vector of effect size estimates at p variants from GWAS trait i and let $\hat{\boldsymbol{\Sigma}}_i$ represent the $p \times p$ diagonal matrix containing squared standard errors. We model the estimated SNP effects for trait i as

$$\hat{\boldsymbol{\beta}}_i \sim \mathcal{N}(\mathbf{L}\mathbf{f}_i + \boldsymbol{\mu}, \boldsymbol{\tau}^{-1}\hat{\boldsymbol{\Sigma}}_i),$$

where $\mathbf{L}$ is a $p \times k$ factor loading matrix shared by all n traits, $\mathbf{f}_i$ is the $k \times 1$ latent factor scores for trait i, $\boldsymbol{\mu}$ is a $p \times 1$ intercept vector. Conditional on shared loadings $\mathbf{L}$ and latent factors $\mathbf{f}_i$, we assume residuals are independent such that the off diagonals of $\hat{\boldsymbol{\Sigma}}_i$ are zeros. In practice, $\hat{\boldsymbol{\beta}}_i$ can be correlated due to linkage disequilibrium (i.e. LD) patterns, however we can perform LD-pruning to analyze a subset of variants such that LD is relatively minimal. Lastly, we let $\boldsymbol{\tau} = \sigma^{-2}$ capture any cross-study heterogeneity.

To simplify our model, we standardize effect sizes by pre-multiplying it with its standard errors such that

$$\hat{\mathbf{z}}_i \sim \mathcal{N}(\hat{\boldsymbol{\Sigma}}_i^{-1/2}(\mathbf{L}\mathbf{f}_i + \boldsymbol{\mu}), \boldsymbol{\tau}^{-1}\mathbf{I}_p).$$

where $\hat{\mathbf{z}}_i = \hat{\boldsymbol{\Sigma}}_i^{-1/2}\hat{\boldsymbol{\beta}}_i$. We note that standard errors for variant j in study i are proportional to $\sqrt{\frac{1}{N_i 2 f_{ij}(1-f_{ij})}}$, where f_{ij} is the minor allele frequency at variant j and N_i is the i^{th} GWAS total sample size. We assume that allele frequencies at variant j are roughly constant across studies when the underlying population reflects similar ancestries. Thus this per-variant scaling term can be absorbed into the loading matrix $\mathbf{L}$ and intercept $\boldsymbol{\mu}$ giving,

$$\hat{\mathbf{z}}_i \sim \mathcal{N}(\sqrt{N_i}(\mathbf{L}\mathbf{f}_i + \boldsymbol{\mu}), \boldsymbol{\tau}^{-1}\mathbf{I}_p).$$

Similar to the Bayesian PCA approach proposed by Bishop 1999 [2], we impose full Bayesian treatment to this factor analysis model. The latent structure modeled by $\mathbf{L} = [\boldsymbol{\ell}^1, \dots, \boldsymbol{\ell}^p]^\intercal$, $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_n]$ and $\boldsymbol{\mu}$ has prior distributions as

$$\begin{aligned}
\Pr(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{0}, \phi^{-1}\mathbf{I}_p) \\
\Pr(\mathbf{F}) &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{f}_i \mid \mathbf{0}, \mathbf{I}_k) \\
\Pr(\mathbf{L} \mid \boldsymbol{\alpha}) &= \prod_{j=1}^{p} \mathcal{N}(\boldsymbol{\ell}^j \mid \mathbf{0}, \operatorname{diag}(\boldsymbol{\alpha}^{-1}))
\end{aligned}$$

To regularize model complexity, we put automatic relevance determination (ARD) priors on the loading matrix $\mathbf{L}$ such that less informative factors are shrunk towards zero. For each factor q, the ARD parameter $\boldsymbol{\alpha}_q$ is proportional to the inverse precision of that factor and modeled as Gamma distribution. The prior distributions are specified as follows:

$$\begin{aligned}
\Pr(\boldsymbol{\alpha} \mid a_\alpha, b_\alpha) &= \prod_{q=1}^{k} \Gamma(\boldsymbol{\alpha}_q \mid a_\alpha, b_\alpha) \\
\Pr(\boldsymbol{\tau} \mid a_\tau, b_\tau) &= \Gamma(\boldsymbol{\tau} \mid a_\tau, b_\tau)
\end{aligned}$$

2.3 Compare to tSVD

Truncated singular value decomposition (tSVD) is a reduced rank representation of original data matrix using result of SVD[3]. The full SVD decomposition for an observed Z-score summary statistics matrix is:

$$\hat{\mathbf{Z}}_{n\times p} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{i=1}^{p} \mathbf{u}_i \mathbf{s}_i \mathbf{v}_i^T$$

Then tSVD has:

$$\hat{\mathbf{Z}}_{n\times p} \approx \sum_{i=1}^{k} \mathbf{u}_i \mathbf{s}_i \mathbf{v}_i^T$$

Unlike model-free tSVD, FactorGo appropriately accounts for the uncertainty in Z-scores due to differential power of GWAS studies and automatically infer model complexity. If N_i is constant across studies and $\boldsymbol{\tau}^{-1}$ approaches 0, then we expect FactorGo produce similar result as tSVD.

2.4 Variational Inference

2.4.1 Overview

Considering the large number of parameters to estimate and its scalability to large dataset, we chose variational inference (VI) over other inference technique such as MCMC to infer posterior distribution of unknown parameters[4]. Unlike MCMC that aims to sample from true posterior distribution, VI converts this estimation problem to optimization problem. Given a choice of a tractable surrogate distribution $Q(\cdot)$ for the non-tractable true posterior, we solve for the optimal estimates to maximize the evidence lower bound (ELBO) of marginal log likelihood of data.

Let $\boldsymbol{\theta} = (\mathbf{L}, \mathbf{F}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ contains all unknown parameters, $\boldsymbol{\eta} = (a_\alpha, b_\alpha, a_\tau, b_\tau, \phi)$ be user specified hyperparameters and outcome Z-score data is $\hat{\mathbf{Z}}$. Suppose $Q(\boldsymbol{\theta})$ is any surrogate distribution for true posteriors, we can show the ELBO is a rigorous lower bound for marginal data likelihood by Janssen's inequality:

$$\begin{aligned}
\log P(\hat{\mathbf{Z}}) &= \log \int P(\hat{\mathbf{Z}}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \log \int Q(\boldsymbol{\theta}) \frac{P(\hat{\mathbf{Z}}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\geq \int Q(\boldsymbol{\theta}) \log \frac{P(\hat{\mathbf{Z}}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= ELBO(Q)
\end{aligned}$$

The difference between $\log P(\hat{\mathbf{Z}})$ and lower bound $ELBO(Q)$ is called Kullback-Leibler (KL) divergence:

$$\begin{aligned}
\log P(\hat{\mathbf{Z}}) - ELBO(Q) &= E_Q(\log P(\hat{\mathbf{Z}})) - ELBO(Q) \\
&= \int Q(\boldsymbol{\theta}) \log P(\hat{\mathbf{Z}}) d\boldsymbol{\theta} - \int Q(\boldsymbol{\theta}) \log \frac{P(\hat{\mathbf{Z}}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int Q(\boldsymbol{\theta}) (\log P(\hat{\mathbf{Z}}) - \log \frac{P(\hat{\mathbf{Z}}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})}) d\boldsymbol{\theta} \\
&= \int Q(\boldsymbol{\theta}) \log \frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta} \mid \hat{\mathbf{Z}})} d\boldsymbol{\theta} \\
&= \mathrm{KL}[Q(\boldsymbol{\theta}) \parallel P(\boldsymbol{\theta} \mid \hat{\mathbf{Z}}, \boldsymbol{\eta})]
\end{aligned}$$

The relationship between these quantities is:

$$ELBO(Q) = \log P(\hat{\mathbf{Z}} \mid \boldsymbol{\eta}) - \mathrm{KL}[Q(\boldsymbol{\theta}) \parallel P(\boldsymbol{\theta} \mid \hat{\mathbf{Z}}, \boldsymbol{\eta})]$$

The *complete-data* likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \hat{\mathbf{Z}}, \boldsymbol{\eta})$ is:

$$\mathcal{L}(\boldsymbol{\theta}, \hat{\mathbf{Z}} \mid \boldsymbol{\eta}) = \Pr(\boldsymbol{\tau} \mid a_\tau, b_\tau) \Pr(\boldsymbol{\alpha} \mid a_\alpha, b_\alpha) \Pr(\mathbf{L} \mid \boldsymbol{\alpha}) \Pr(\boldsymbol{\mu} \mid \phi) \prod_n \Pr(\hat{\mathbf{z}}_n \mid \boldsymbol{\mu}, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\alpha}, \boldsymbol{\tau}, N_i) \Pr(\mathbf{f}_i)$$

Here we chose a fully factorizable distribution Q from conjugate family such that:

$$Q(\mathbf{L}, \mathbf{F}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\mu}) = Q(\mathbf{L}) Q(\mathbf{F}) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\tau}) Q(\boldsymbol{\mu})$$

The solution for each parameter $\boldsymbol{\theta}_i$ is found by maximizing the the lower bound with respective to the following quantity:

$$Q(\boldsymbol{\theta}_i) \propto \mathbb{E}_{-i}[\log \mathcal{L}(\hat{\mathbf{Z}}, \boldsymbol{\theta})]$$

where the expectation is with respect to all parameters except $\boldsymbol{\theta}_i$.

Using completing squares, we can write out the posterior means and variances for multivariate Gaussian distribution or Gamma distribution. Here we provide solutions for each Q.

2.4.2 $Q(\mathbf{F})$

$$
\begin{aligned}
Q(\mathbf{F}) &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{f}_i \mid \mathbf{m}_{\mathbf{f}_i}, \mathbf{V}_{\mathbf{f}_i}) \text{ where} \\
\mathbf{V}_{\mathbf{f}_i} &= (\mathbf{I}_k + N_i \mathbb{E}[\boldsymbol{\tau}]\mathbb{E}[\mathbf{L}^\intercal \mathbf{L}])^{-1} \\
&= \left(\mathbf{I}_k + N_i \mathbb{E}[\boldsymbol{\tau}] \sum_{j=1}^{p} \mathbb{E}[\boldsymbol{\ell}^j \boldsymbol{\ell}^{j\intercal}]\right)^{-1} \\
&= \left(\mathbf{I}_K + N_i \mathbb{E}[\boldsymbol{\tau}] \sum_{j=1}^{p} (\mathbf{V}_{\boldsymbol{\ell}^j} + \mathbf{m}_{\boldsymbol{\ell}^j} \mathbf{m}_{\boldsymbol{\ell}^j}^\intercal)\right)^{-1} \\
\mathbf{m}_{\mathbf{f}_i} &= \sqrt{N_i}\mathbb{E}[\boldsymbol{\tau}]\mathbf{V}_{\mathbf{f}_i}\mathbb{E}[\mathbf{L}]^\intercal(\hat{\mathbf{z}}_i - \sqrt{N_i}\mathbb{E}[\boldsymbol{\mu}])
\end{aligned}
$$

2.4.3 $Q(\mu)$

$$
\begin{aligned}
Q(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{m}_{\boldsymbol{\mu}}, \mathbf{V}_{\boldsymbol{\mu}}) \text{ where} \\
\mathbf{V}_{\boldsymbol{\mu}} &= (\phi + \mathbb{E}[\boldsymbol{\tau}] \sum_{i=1}^{n} N_i)^{-1} \mathbf{I}_p \\
\mathbf{m}_{\boldsymbol{\mu}} &= \mathbb{E}[\boldsymbol{\tau}]\mathbf{V}_{\boldsymbol{\mu}} \sum_{i=1}^{n} \sqrt{N_i}(\hat{\mathbf{z}}_i - \sqrt{N_i}\mathbb{E}[\mathbf{L}]\mathbb{E}[\mathbf{f}_i])
\end{aligned}
$$

2.4.4 $Q(\mathbf{L})$

$$
\begin{aligned}
Q(\mathbf{L}) &= \prod_{j=1}^{p} \mathcal{N}(\boldsymbol{\ell}^j \mid \mathbf{m}_{\boldsymbol{\ell}^j}, \mathbf{V}_{\boldsymbol{\ell}^j}) \text{ where} \\
\mathbf{V}_{\boldsymbol{\ell}^j} &= \left(\operatorname{diag}(\mathbb{E}[\boldsymbol{\alpha}]) + \mathbb{E}[\boldsymbol{\tau}] \sum_{i=1}^{n} N_i \mathbb{E}[\mathbf{f}_i \mathbf{f}_i^\intercal]\right)^{-1} \\
&= \left(\operatorname{diag}(\mathbb{E}[\boldsymbol{\alpha}]) + \mathbb{E}[\boldsymbol{\tau}] \sum_{i=1}^{n} N_i (\mathbf{V}_{\mathbf{f}_i} + \mathbf{m}_{\mathbf{f}_i} \mathbf{m}_{\mathbf{f}_i}^\intercal)\right)^{-1} \\
\mathbf{m}_{\boldsymbol{\ell}^j} &= \mathbb{E}[\boldsymbol{\tau}]\mathbf{V}_{\boldsymbol{\ell}^j} \left(\sum_{i=1}^{n} \sqrt{N_i}\mathbb{E}[\mathbf{f}_i](\hat{\mathbf{z}}_{ij} - \sqrt{N_i}\mathbb{E}[\boldsymbol{\mu}_j])\right)
\end{aligned}
$$

2.4.5 $Q(\alpha)$

$$
\begin{aligned}
Q(\boldsymbol{\alpha}) &= \prod_k \Gamma(\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_{\alpha k}) \text{ where} \\
\tilde{a}_\alpha &= a_\alpha + \frac{p}{2} \\
\tilde{b}_{\alpha k} &= b_\alpha + \frac{\mathbb{E}[\boldsymbol{\ell}_k^\intercal \boldsymbol{\ell}_k]}{2} = b_\alpha + \frac{\operatorname{diag}(\mathbb{E}[\mathbf{L}^\intercal \mathbf{L}])}{2} \\
&= b_\alpha + \frac{1}{2}\sum_{j=1}^{p} \mathbb{E}[\boldsymbol{\ell}^j \boldsymbol{\ell}^{j\intercal}]_{(kk)} \\
&= b_\alpha + \frac{1}{2}\sum_{j=1}^{p} [\mathbf{V}_{\boldsymbol{\ell}^j} + \mathbf{m}_{\boldsymbol{\ell}^j} \mathbf{m}_{\boldsymbol{\ell}^j}^\intercal]_{(kk)}
\end{aligned}
$$

2.4.6 $Q(\tau)$

$$
\begin{aligned}
Q(\boldsymbol{\tau}) &= \Gamma(\boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau) \text{ where} \\
\tilde{a}_\tau &= a_\tau + \frac{np}{2} \\
\tilde{b}_\tau &= b_\tau + \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\|\hat{\mathbf{z}}_i - \sqrt{N_i}\mathbf{L}\mathbf{f}_i - \sqrt{N_i}\boldsymbol{\mu}\|^2
\end{aligned}
$$

2.4.7 ELBO

After each iteration, calculate ELBO by $\mathbb{E}[\log \Pr(\hat{\mathbf{Z}} \mid \boldsymbol{\theta}, \boldsymbol{\eta})] - \mathrm{KL}[Q(\boldsymbol{\theta}) \parallel \Pr(\boldsymbol{\theta} \mid \boldsymbol{\eta})]$, where KL term can be calculate separately for each parameter.

Data likelihood term:

$$
\mathbb{E}[\log \Pr(\hat{\mathbf{Z}} \mid \boldsymbol{\theta}, \boldsymbol{\eta})] = \mathbb{E}\left[\log \prod_{i=1}^{n} \mathcal{N}(\sqrt{N_i}\mathbf{L}\mathbf{f}_i + \sqrt{N_i}\boldsymbol{\mu}, \boldsymbol{\tau}^{-1}) \mid \boldsymbol{\theta}, \boldsymbol{\eta}\right]
$$

L:

$$
\begin{aligned}
\mathbb{E}[\log Q(\boldsymbol{\ell}^j)] &= \mathbb{E}\left[-\frac{k}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}_{\boldsymbol{\ell}^j}| - \frac{1}{2}(\boldsymbol{\ell}^j - \mathbf{m}_{\boldsymbol{\ell}^j})^\intercal \mathbf{V}_{\boldsymbol{\ell}^j}^{-1}(\boldsymbol{\ell}^j - \mathbf{m}_{\boldsymbol{\ell}^j}) \mid \boldsymbol{\theta}\right] \\
\mathbb{E}[\log \Pr(\boldsymbol{\ell}^j \mid \boldsymbol{\eta})] &= \mathbb{E}\left[-\frac{k}{2}\log 2\pi - \frac{1}{2}\log|\operatorname{diag}(\boldsymbol{\alpha}^{-1})| - \frac{1}{2}(\boldsymbol{\ell}^{j\intercal}\operatorname{diag}(\boldsymbol{\alpha})\boldsymbol{\ell}^j) \mid \boldsymbol{\eta}\right] \\
\mathrm{KL}[Q(\mathbf{L}) \parallel \Pr(\mathbf{L} \mid \boldsymbol{\eta})] &= \sum_{j=1}^{p} \mathbb{E}[\log Q(\boldsymbol{\ell}^j)] - \mathbb{E}[\log \Pr(\boldsymbol{\ell}^j \mid \boldsymbol{\eta})]
\end{aligned}
$$

F:

$$
\mathrm{KL}[Q(\mathbf{F}) \parallel \Pr(\mathbf{F} \mid \boldsymbol{\eta})] = \sum_{i=1}^{n} \mathbb{E}[\log Q(\mathbf{f}_i)] - \mathbb{E}[\log \Pr(\mathbf{f}_i \mid \boldsymbol{\eta})]
$$

$\boldsymbol{\mu}$:

$$\mathrm{KL}[Q(\boldsymbol{\mu}) \parallel \Pr(\boldsymbol{\mu} \mid \boldsymbol{\eta})] = \mathbb{E}[\log Q(\boldsymbol{\mu})] - \mathbb{E}[\log \Pr(\boldsymbol{\mu} \mid \boldsymbol{\eta})]$$

$\boldsymbol{\alpha}$:

$$\begin{aligned}
\mathbb{E}[\log \Gamma(\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha)] &= \tilde{a}_\alpha \log(\tilde{b}_\alpha) + (\tilde{a}_\alpha - 1)\mathbb{E}[\log \boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha] - \tilde{b}_\alpha \mathbb{E}[\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha] - \log \Gamma(\tilde{a}_\alpha) \\
\mathbb{E}[\log \Gamma(\boldsymbol{\alpha}_k \mid a_\alpha, b_\alpha)] &= a_\alpha \log(b_\alpha) + (a_\alpha - 1)\mathbb{E}[\log \boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha] - b_\alpha \mathbb{E}[\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha] - \log \Gamma(a_\alpha) \\
\mathrm{KL}[Q(\boldsymbol{\alpha}) \parallel \Pr(\boldsymbol{\alpha} \mid \boldsymbol{\eta})] &= \sum_k \mathbb{E}[\log \Gamma(\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, \tilde{b}_\alpha)] - \mathbb{E}[\log \Gamma(\boldsymbol{\alpha}_k \mid a_\alpha, b_\alpha)]
\end{aligned}$$

$\boldsymbol{\tau}$:

$$\begin{aligned}
\mathbb{E}[\log \Gamma(\boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau)] &= \tilde{a}_\tau \log(\tilde{b}_\tau) + (\tilde{a}_\tau - 1)\mathbb{E}[\log \boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau] - \tilde{b}_\tau \mathbb{E}[\boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau] - \log \Gamma(\tilde{a}_\tau) \\
\mathbb{E}[\log \Gamma(\boldsymbol{\tau} \mid a_\tau, b_\tau)] &= a_\tau \log(b_\tau) + (a_\tau - 1)\mathbb{E}[\log \boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau] - b_\tau \mathbb{E}[\boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau] - \log \Gamma(a_\tau) \\
\mathrm{KL}[Q(\boldsymbol{\tau}) \parallel \Pr(\tau \mid \boldsymbol{\eta})] &= \mathbb{E}[\log \Gamma(\boldsymbol{\tau} \mid \tilde{a}_\tau, \tilde{b}_\tau)] - \mathbb{E}[\log \Gamma(\boldsymbol{\tau} \mid a_\tau, b_\tau)]
\end{aligned}$$

2.5 Parameter expansion

The convergence of FactorGo under the above model can be slow because **L** and **F** are strongly coupled in the model, whereas vectors in **L** and **F** are assumed to be independent a posteriori for computational convenience. To speed up inference, we applied a parameter expansion method proposed for variational Bayesian factor analysis specifically[5]. The general idea is to introduce auxilliary parameter for bias **b** and **R** in the posterior distribution that are optimized during inference. After each iteration step, we can jointly update the parameters in $\mathbf{L}, \mathbf{F}, \boldsymbol{\alpha}$. Here we provided the updating rule under this transformation method:

1. First remove bias between **F** and $\boldsymbol{\mu}$

$$\begin{aligned}
Q^*(\mathbf{F}) &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{f}_i \mid \mathbf{m}_{\mathbf{f}_i} - \mathbf{b}, \mathbf{V}_{\mathbf{f}_i}) \\
Q^*(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{m}_{\boldsymbol{\mu}} + \mathbf{L}\mathbf{b}, \mathbf{V}_{\boldsymbol{\mu}})
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{b} &= (\sum_{i=1}^{n} \Psi_i)^{-1} (\sum_{i=1}^{n} \Psi_i \mathbf{m}_{\mathbf{f}_i}) \\
\Psi_i &= N_i \mathbb{E}[\boldsymbol{\tau}] p \mathbf{V}_{\boldsymbol{\ell}^j} + \mathbf{I}_k
\end{aligned}$$

2. Then rotate latent subspace

$$\begin{aligned}
Q^*(\mathbf{L}) &= \prod_{j=1}^{p} \mathcal{N}(\boldsymbol{\ell}^j \mid \mathbf{R}^\intercal \mathbf{m}_{\boldsymbol{\ell}^j}, \mathbf{R}^\intercal \mathbf{V}_{\boldsymbol{\ell}^j} \mathbf{R}) \\
Q^*(\mathbf{F}) &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{f}_i \mid \mathbf{R}^{-1} \mathbf{m}_{\mathbf{f}_i}, \mathbf{R}^{-1} \mathbf{V}_{\mathbf{f}_i} \mathbf{R}^{-\intercal}) \\
Q^*(\boldsymbol{\alpha}) &= \prod_{k} \Gamma(\boldsymbol{\alpha}_k \mid \tilde{a}_\alpha, b_\alpha + \frac{1}{2} \mathrm{diag}(\mathbf{R}^\intercal \mathbb{E}_Q[\mathbf{L}^\intercal \mathbf{L}] \mathbf{R}))
\end{aligned}$$

The optimal rotation matrix $\mathbf{R}$ can be found by following steps:
let $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$, then $\mathbf{U}$ and $\mathbf{\Lambda}$ are found by eigen-decomposition of:

$$\frac{1}{n}\mathbb{E}_Q[\mathbf{F}\mathbf{F}^\intercal] = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\intercal$$

$\mathbf{V}$ is found by eigen-decomposition of:

$$\mathbf{\Lambda}\mathbf{U}^\intercal\mathbb{E}_Q[\mathbf{L}^\intercal\mathbf{L}]\mathbf{U}\mathbf{\Lambda} = \mathbf{V}\mathbf{D}\mathbf{V}^\intercal$$

2.5.1 FactorGo algorithm

Algorithm 1: FactorGo with parameter expansion design

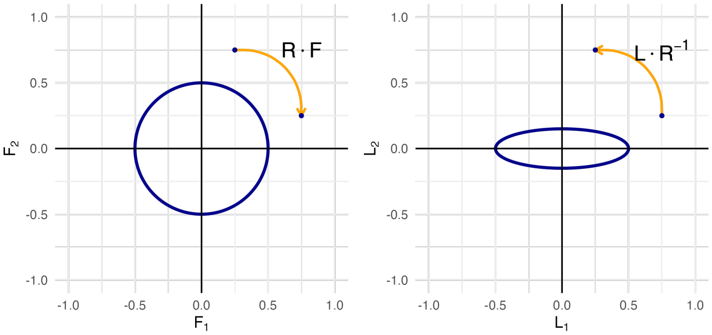
```
Input: GWAS Z-score summary data and sample size
Initialize: $a_\alpha = b_\alpha = a_\tau = b_\tau = \phi = 10^{-5}, \mathbb{E}[\boldsymbol{\ell}^j] = \mathbf{0}, \mathbb{V}(\boldsymbol{\ell}^j) = \mathbf{I}$ for all $j \in [p], ELBO_0 = 0$
while $ELBO_i - ELBO_{i-1} > 0.001$ or $i \leq itr_{max}$ do
    update $\mathbf{m}_\mathbf{F}, \mathbf{V}_\mathbf{F}$
    update $\mathbf{m}_{\boldsymbol{\mu}}, \mathbf{V}_{\boldsymbol{\mu}}$
    update $\mathbf{m}_\mathbf{L}, \mathbf{V}_\mathbf{L}$
    update $\mathbf{m}_{\boldsymbol{\alpha}}, \mathbf{V}_{\boldsymbol{\alpha}}$
    find optimal $\mathbf{b}, \mathbf{R}$ for transformation and update above parameters
    update $\mathbf{m}_{\boldsymbol{\tau}}, \mathbf{V}_{\boldsymbol{\tau}}$
    Calculate $ELBO$
end
Output: return posterior mean and variance of $\boldsymbol{\theta} = (\mathbf{L}, \mathbf{F}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\tau})$.
```

2.6 FactorGo model identifiablility

The rotation of $\mathbf{F}$ and $\mathbf{L}$ will not change the data likelihood $\mathcal{L}(\hat{\mathbf{Z}} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ because

$$\hat{\mathbf{Z}}_i = \sqrt{N_i}(\mathbf{L}\mathbf{F}_i + \boldsymbol{\mu}) + \boldsymbol{\epsilon}_i = \sqrt{N_i}((\mathbf{L}\mathbf{R}^{-1})(\mathbf{R}\mathbf{F}_i) + \boldsymbol{\mu}) + \boldsymbol{\epsilon}_i$$

where $\mathbf{R}$ is orthogonal rotational matrix such that $\mathbf{R}^{-1}\mathbf{R} = \mathbf{I}$. However, this rotation will change the complete-data likelihood $\mathcal{L}(\hat{\mathbf{Z}}, \boldsymbol{\theta} \mid \boldsymbol{\eta})$ because of ARD priors on L. The ARD prior creates different scales on factor loading along each axis as shown below:



Example of rotation in F and L. The shape of $\mathbf{L}$ for two factors is an ellipse, where each factor axis has a different scale constructed by ARD priors.

2.7 Calculate variance explained R^2 in observed data

Here we provide formula to calculate variance explained by each factor R_k^2. For k^{th} factor, and i^{th} study, let $\hat{\mathbf{z}}_i'$ be fitted Z-score value by all factors and $\hat{\mathbf{z}}_i^{(k)'}$ be fitted value for k^{th} factor only:

$$\begin{aligned}\hat{\mathbf{z}}_i' &= \sqrt{N_i}\mathbb{E}[\mathbf{L}]\mathbb{E}[\mathbf{f}_i] = \sqrt{N_i}\sum_k \mathbb{E}[\mathbf{f}_{ik}]\mathbb{E}[\boldsymbol{\ell}_k] \\ \hat{\mathbf{z}}_i^{(k)'} &= \sqrt{N_i}\mathbb{E}[\mathbf{f}_{ik}]\mathbb{E}[\boldsymbol{\ell}_k]\end{aligned}$$

Then calculate R_k^2 using total residual error and total variance, where σ^2 is canceled:

$$\begin{aligned}SSE_k &= \sum_i (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_i^{(k)'})^\intercal \sigma^{-2} (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_i^{(k)'}) \\ TSS &= \sum_i \hat{\mathbf{z}}_i^\intercal \sigma^{-2} \hat{\mathbf{z}}_i \\ R_k^2 &= 1 - \frac{SSE_k}{TSS}\end{aligned}$$

Supplemental Note 2

Characterizing second and third leading factors for focal traits in FactorGo

BMI

The next two leading factors for BMI identified its shared biology with pharynx and digestive system respectively (squared cosine score: 17.66%, 2.86%). Factor 4 characterized standing height (2.88%) and body fat measures (cumulative 22.73% across 13 traits; **Figure S24, Table S2**). Two leading variants were proximal to *PDE10A* (rs9459529:T>C: 0.030%) and *IGF1* (rs1113483:T>C: 0.029%), both of which play roles in energy homeostasis and obesity etiology[6,7] . Factor 4 was enriched with genes specifically expressed in pharynx, which could reflect the negative association between upper airway size and body fat distribution[8]. Factor 7 characterized blood pressure traits that are strongly associated with BMI (25.12% across 25 traits; **Figure S24**)[9]. It was enriched with subcutaneous fat and digestive systems such as the colon and intestines, which supports obesity as a risk factor for colorectal cancer[10].

Height

The next two leading factors for standing height identified its shared biology with cardiovascular and immunity respectively (squared cosine score: 9.64%, 8.51%). Driven by variant proximal to *TBX20* associated with heart growth (rs702843:G>C: 0.059%)[11], factor 6 exhibited enrichment in coronary arteries (**Figure S27**), which is consistent with previous findings of enrichment in coronary tissues in height associated variants[12,13]. Factor 11 was driven by variants closest to *CCL27* and *MBL2* (rs2812349:T>C: 0.036%; rs189269936:C>G: 0.036%), both of which are involved in the immune system[14,15]. This was supported by its enrichment in nasal mucosa harboring diverse immune cells. Although the relationship between the immune system and human growth is unclear, it has been shown there is an energetic tradeoff between immune function and growth in the Amazonian[16].

Rheumatoid arthritis

The next two leading factors for RA identified its shared biology with kidney, liver and urinary bladder (squared cosine score: 6.01%, 5.75%). Factor 75 characterized alkaline phosphatase (2.54%), cardiac (1.81%) and bladder disease (1.16%; **Figure S30**). The leading variants were close to *ITGA9* (rs73055093:C>T: 0.042%) and *IL12B* (rs113630578:A>G: 0.041%), both of which are important for the immune system. This factor showed enrichment in the kidney cortex and liver, both of which contain high levels of ALP enzymes. This can reflect the comorbidity of RA involved with liver and kidney disease[17,18]. Factor 76 characterized traits involved in the central nervous system such as intervertebral disc disorders (1.28%) and peripheral disorders such as foot deformities (0.97%; **Figure S30**). Its enrichment in the urinary bladder could reflect the shared symptoms of general reactive arthritis.

Prostate cancer

The next two leading factors for PCa identified BMI and hormonal disorder associated with prostate cancer respectively (squared cosine score: 7.71%, 6.37%). Since factor 2 was identified as BMI factor (see **Result**), this supported the impact of obesity on prostate cancer progression due to inflammation and metabolic mechanisms[19,20]. Driven by *FOXE1* associated with thyroid morphogenesis (**Figure S33**)[21], factor 58 characterized blood-related traits such as albumin/globulin ratio (2.53%), platelet crit (1.78%) and other hormonal disorders such as hypothyroidism (1.26%), suggesting the shared mechanisms of PCa involved with hormones.

Supplemental Figures

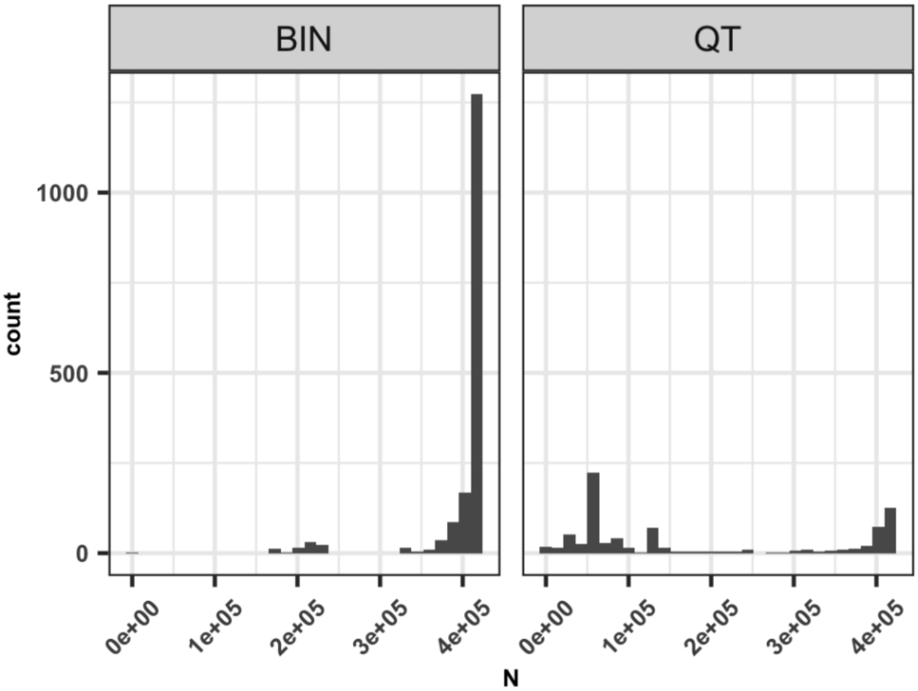


Figure S1. GWAS sample size distribution in Pan-UK Biobank.
Histogram of GWAS total sample size of 2,483 studies from Pan-UK Biobank Europeans (max N=420,531) by 1,677 binary (BIN) traits and 806 quantitative (QT) traits. In simulation, the GWAS sample size was sampled empirically from this distribution.

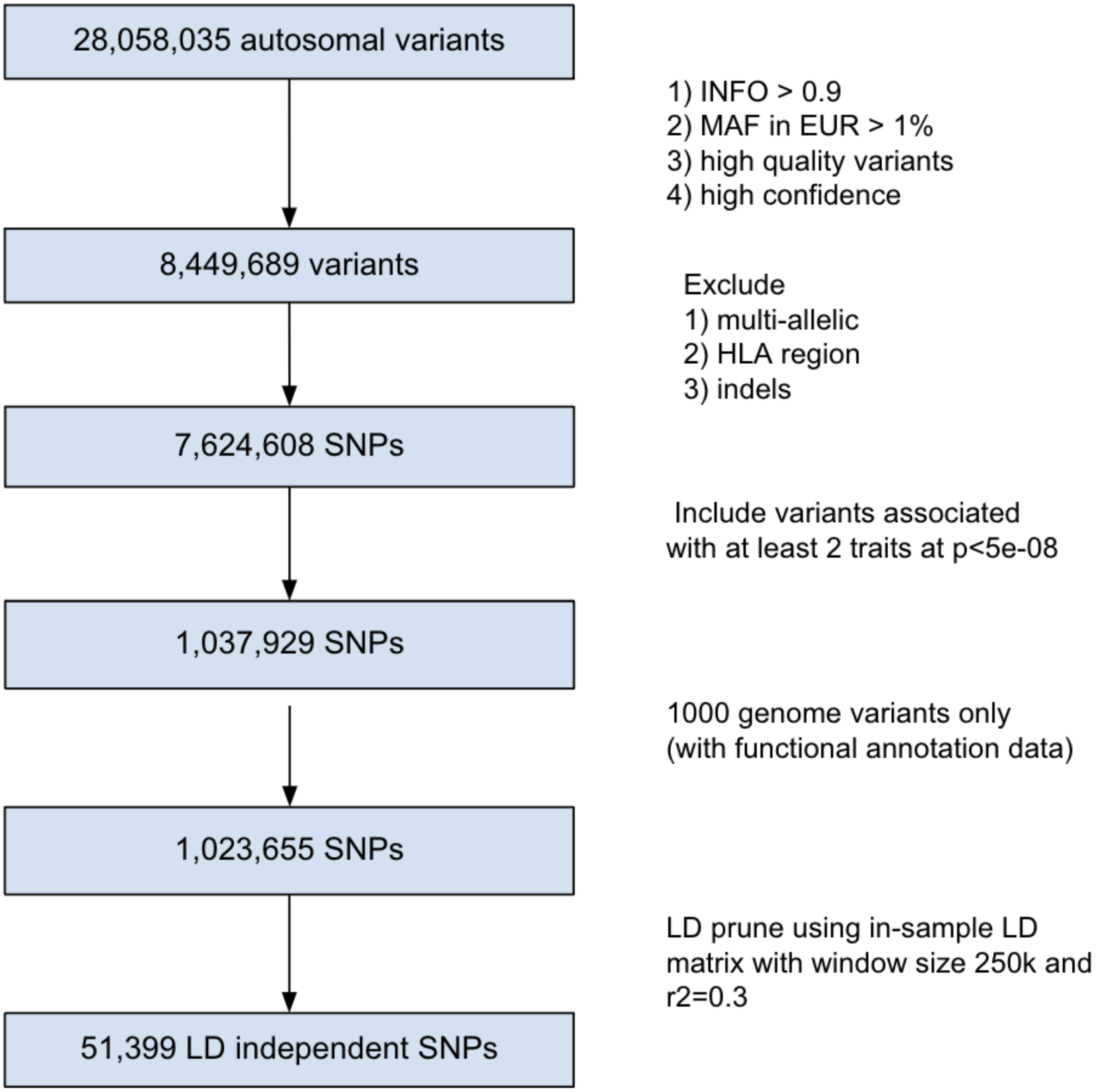

28,058,035 autosomal variants
1) INFO > 0.9
2) MAF in EUR > 1%
3) high quality variants
4) high confidence
8,449,689 variants
Exclude
1) multi-allelic
2) HLA region
3) indels
7,624,608 SNPs
Include variants associated
with at least 2 traits at p<5e-08
1,037,929 SNPs
1000 genome variants only
(with functional annotation data)
1,023,655 SNPs
LD prune using in-sample LD
matrix with window size 250k and
r2=0.3
51,399 LD independent SNPs

Figure S2. Summary of genetic variants filtering in real data analysis of Pan-UK Biobank.

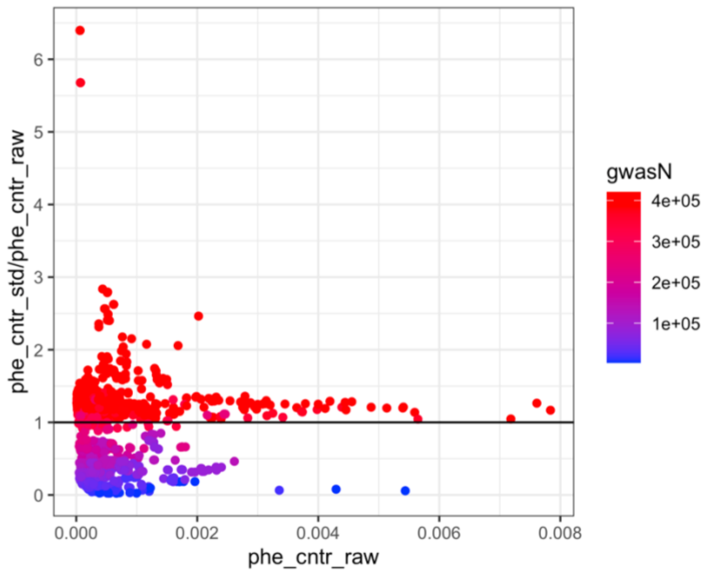


Figure S3. The division of posterior variance in trait contribution score upweight traits with greater sample size.

For each trait on the plot, we plotted the ratio of its mean contribution score across factors versus non-adjusted scores (raw). Since traits will larger GWAS sample size tend to have less uncertainty, so that the division of their posterior variance will upweight trait will larger sample size.

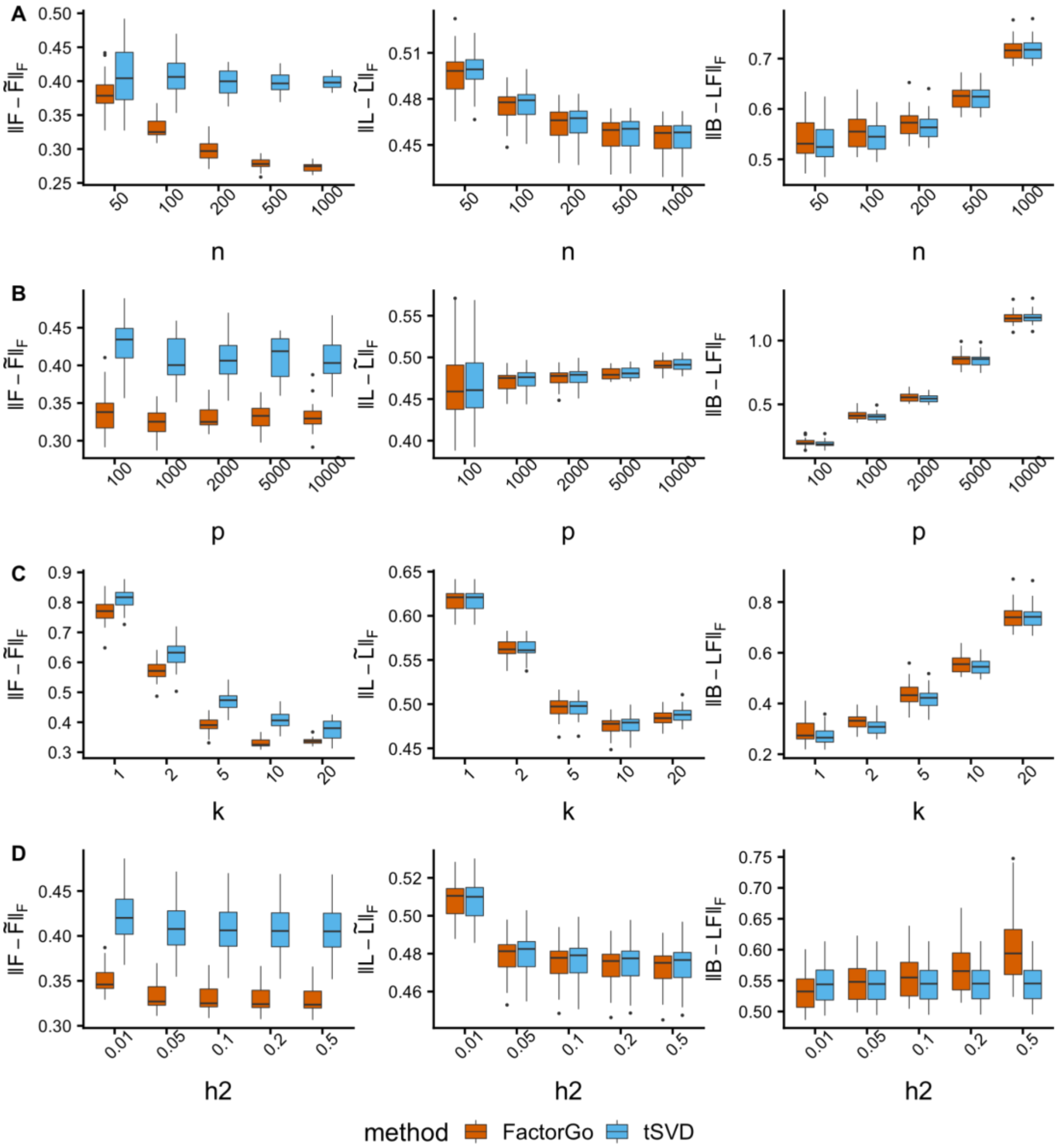


Figure S4. FactorGo outperforms tSVD in trait factor scores in each scenario under model assumptions.

Trait factor score error $||F - \tilde{F}\ ||_F$, variant loading error $||L - \tilde{L}\ ||_F$ and Genetic effect error $||B - LF||_F$ under 4 varying parameters: **(A)** number of studies (n), fix $p = 2000$, $k = 10$, $h^2{}_g = 0.1$; **(B)** number of SNPs (p), fix $p = 2000$, $n = 100$, $h^2{}_g = 0.1$; **(C)** number of true latent factors (k), fix $p = 2000$, $n = 100$, $h^2{}_g = 0.1$; **(D)** SNP heritability ($h^2{}_g$), fix $p = 2000$, $n = 100$, $k = 10$.

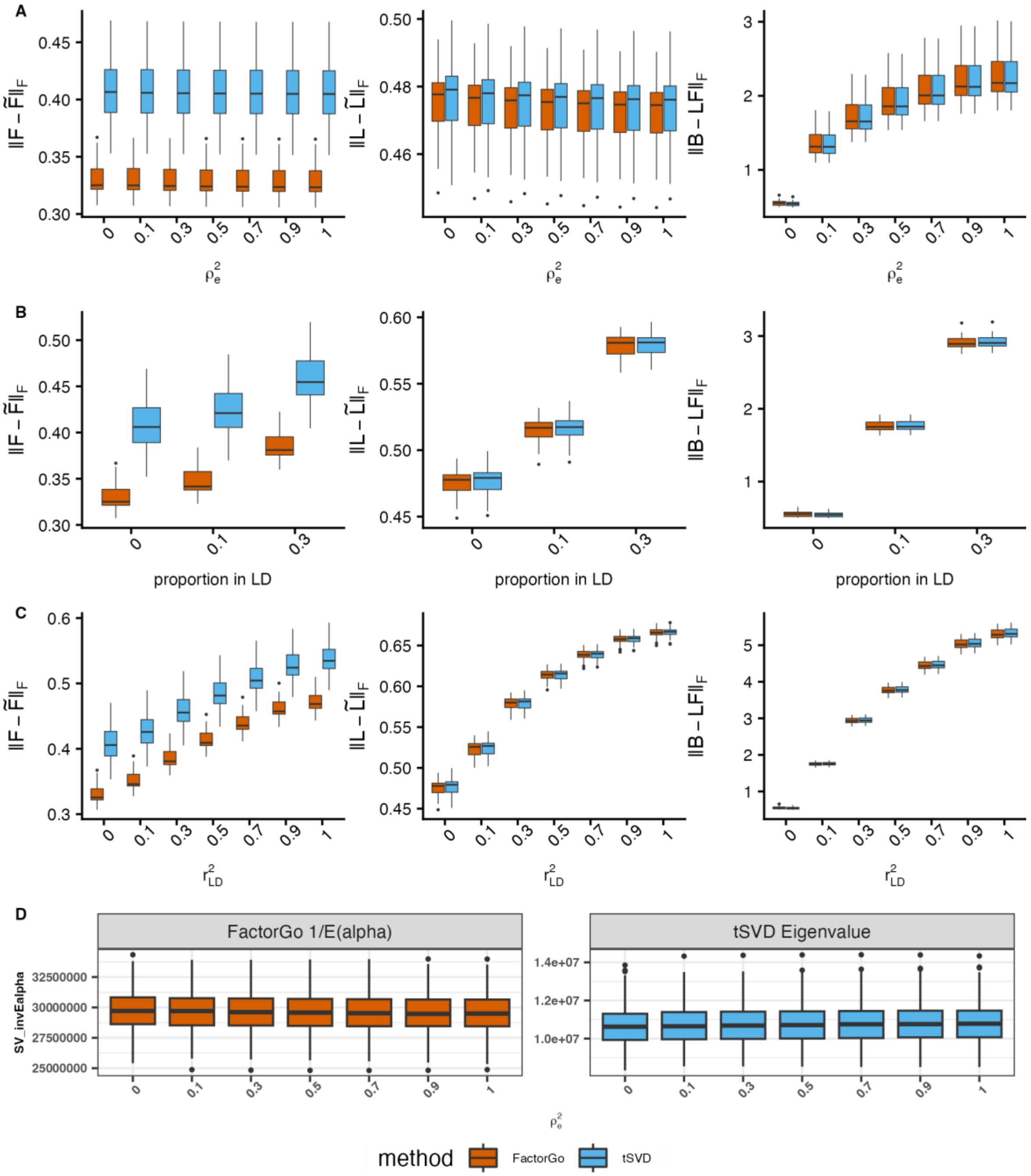


Figure S5. FactorGo is robust to correlated standard errors and outperforms tSVD in learning trait factor scores.

Fix $p = 2000$, $n = 100$, $k = 10$, ${h^2}_g = 0.1$, we simulated **(A)** correlated standard errors with varying residual correlation coefficient ${\rho^2}_e$; adjacent pairs of SNPs in LD with **(B)** varying proportion of SNPs in LD at fixed ${r^2}_{LD} = 0.3$ or **(C)** different magnitude of correlation at fixed proportion 30%. When there are no latent structures (i.e., no pleiotropy) and only correlated errors **(D)**, we ran both methods with $k = 10$ and plotted the $1/E(\alpha)$ from FactorGo and eigenvalues from tSVD methods by correlation magnitude ${\rho^2}_e$.

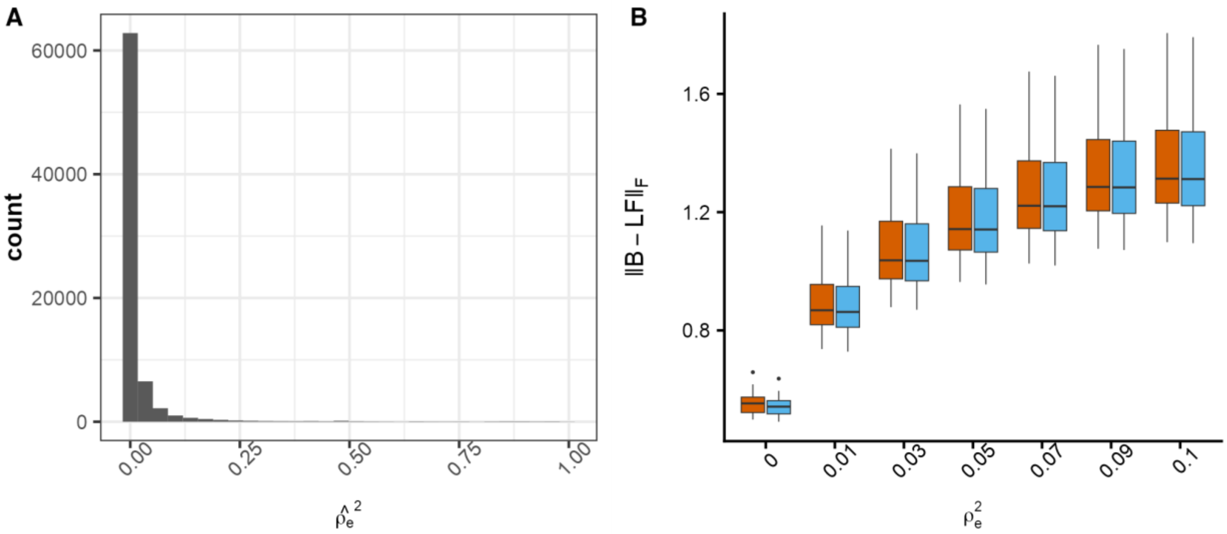


Figure S6. Histogram of residual correlation estimated in all pairwise of 278 Pan-UK Biobank traits.

(A). For 278 highly heritable traits (h2 Z score > 6), we estimated the residual correlation by running all pairwise genetic correlation analysis using LDSC and plotted them. The average squared residual correlation is 0.057 (SD=0.25). (B) Genetic effect error $||B - LF||_F$ at varying residual correlation from 0 to 0.1.

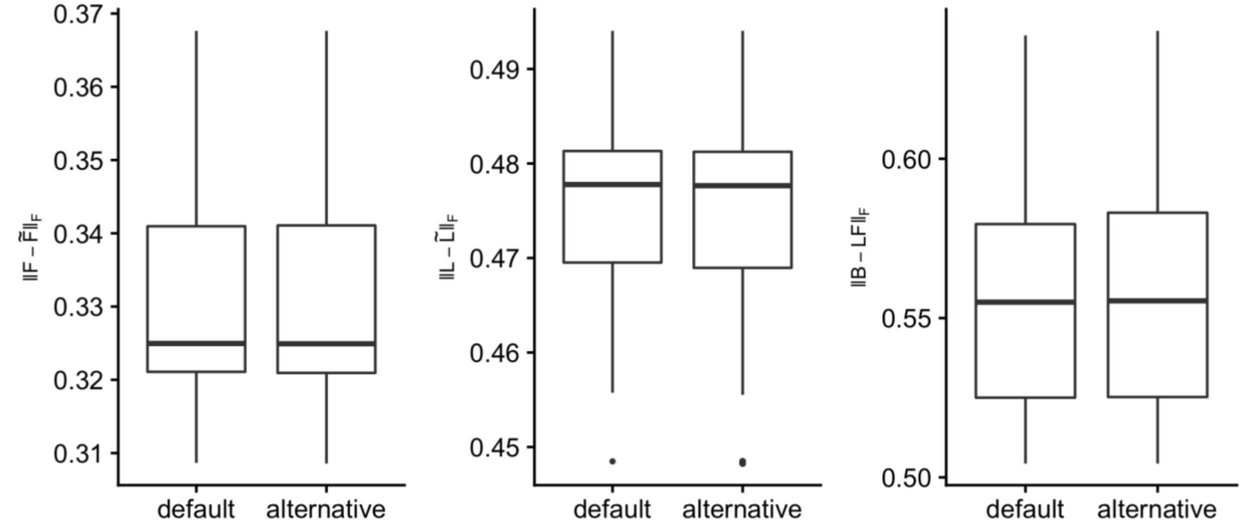


Figure S7. FactorGo is robust to different choices of hyperparameters in simulations. Trait factor score error $||F - \tilde{F}\ ||_F$, variant loading error $||L - \tilde{L}\ ||_F$ and Genetic effect error $||B - LF||_F$ by hyperparameter specification. For each of the 30 simulated data when fixing $p = 2000$, $n = 100$, $k = 10$, $h^2{}_g = 0.1$, we ran FactorGo using all combinations of 1E-05 and 1E-03 for five hyperparameters (total 2^5). The "default" is 1E-05 for all five hyperparameters. Here we compared the reconstruction errors under default setting versus all other alternative settings.

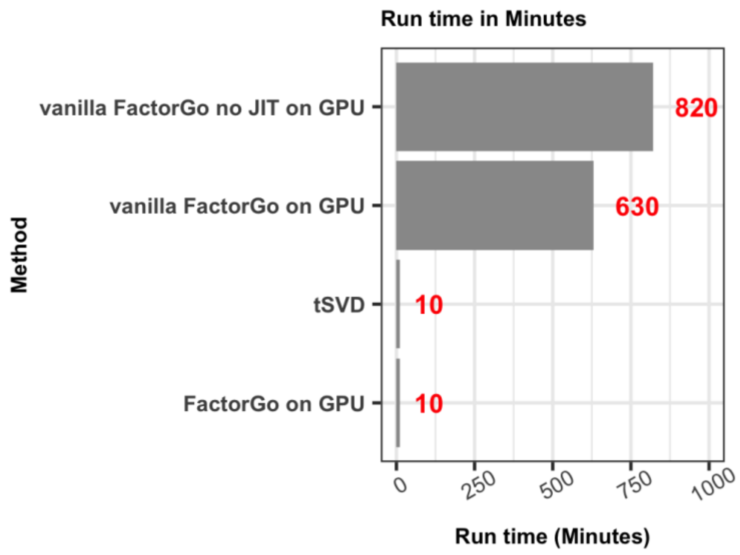


Figure S8. FactorGo and tSVD have the same run time for real data analysis of 2,483 traits and 51,399 variants.

In contrast to vanilla FactorGo, the default FactorGo implements a parameter expansion design (**Note S1**). JIT (Just-In-Time) is a fast execution of python code through the *JAX* package (**Web resources**). Vanilla FactorGo: implementing FactorGo model without speed improvement by parameter expansion design, JIT and GPU.

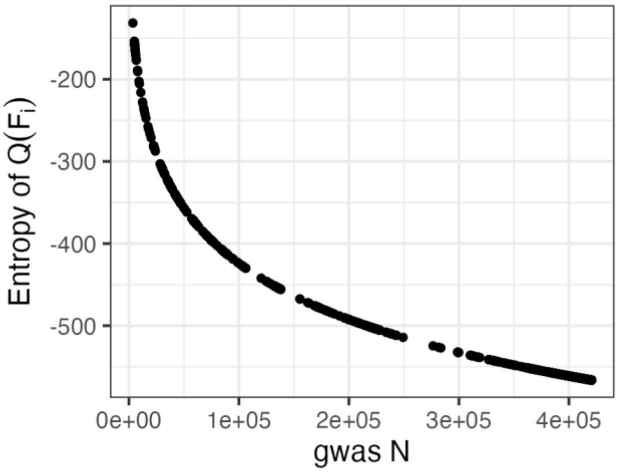


Figure S9. Traits with greater sample size produces less uncertainty in posterior estimates.

The entropy of posterior covariance estimates for factor scores in each trait quantifies uncertainty in posterior mean inference. As expected, traits with greater sample size provides more certainty in posterior inference and thus have lower entropy values in factor score posterior covariance matrix.

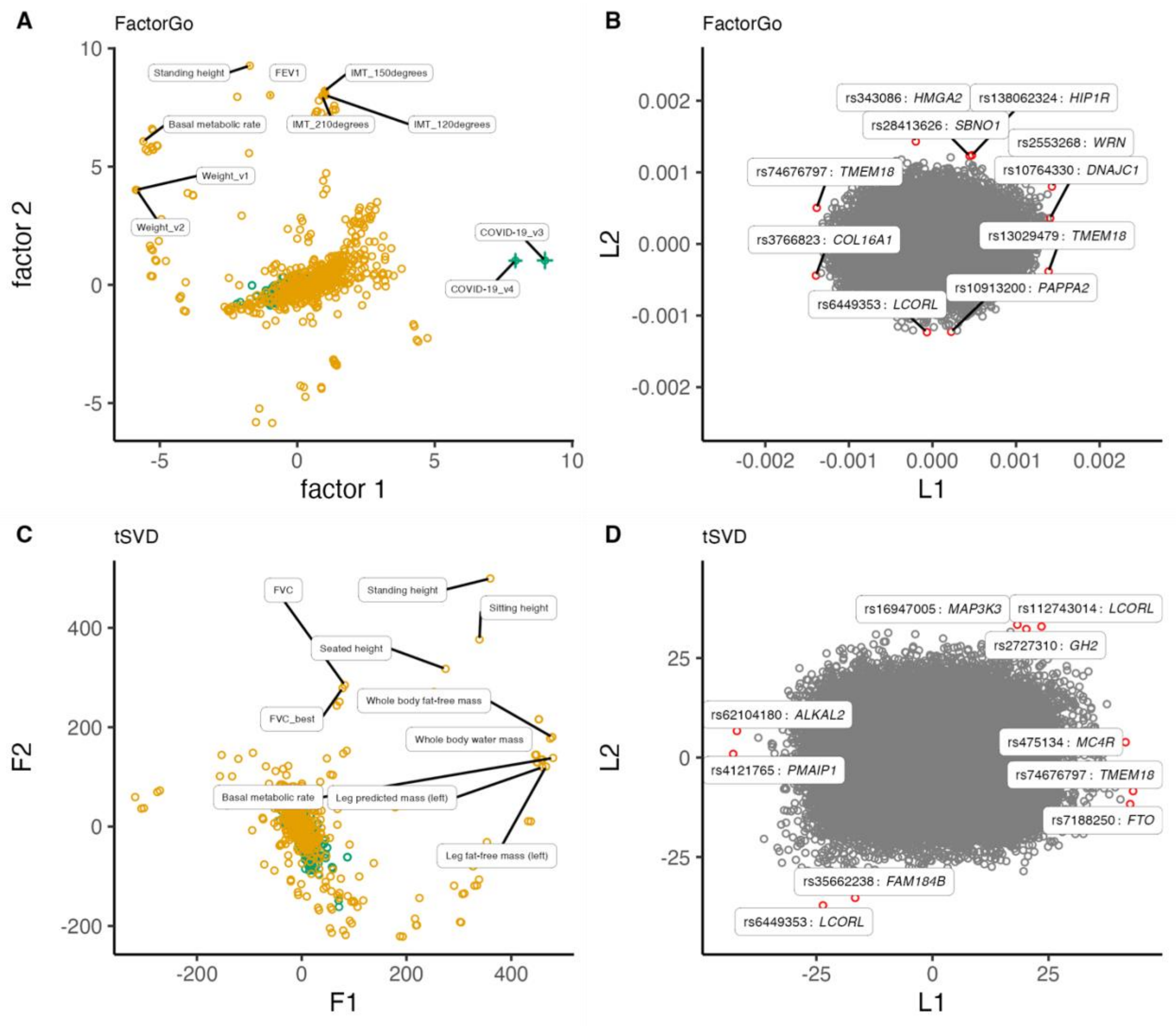


Figure S10. Trait factor scores and variant loading scores for the top two factors in FactorGo and tSVD.

We highlighted 10 leading traits and 10 leading variants (red) for top two factors from FactorGo and tSVD. F1 and F2 are factor scores. L1 and L2 are variant loadings. Binary and quantitative traits are colored differently. Weight_v1: amalgamated measure of weight by multiple means; Weight_v2: weight measured during impedance measurement; FEV1: Forced expiratory volume in 1-second; IMT: Mean carotid IMT (intima-medial thickness); FVC_best: Forced vital capacity, best measure.

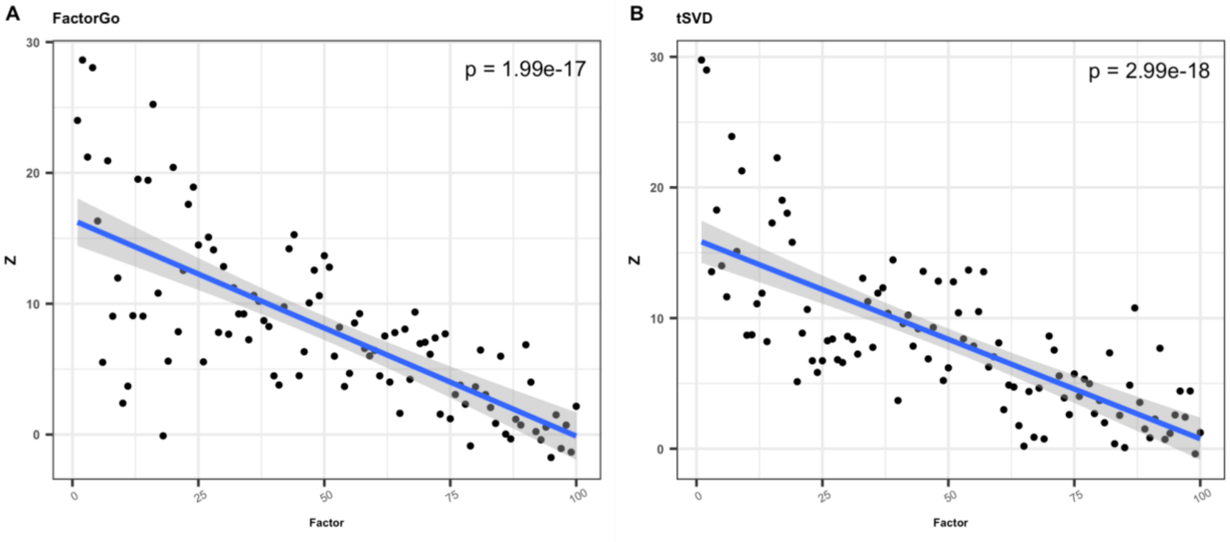


Figure S11. Factor scores in top factors are driven by traits with high heritability.
For **(A)** FactorGo and **(B)** tSVD respectively, each point is the test statistics Z-score for linear association between trait factor scores and the observed heritability estimated by LDSC for 2,305 traits with heritability estimates. Blue line is the fitted regression line with gray confidence band over these 100 points on each plot. This association decay with factor rank.

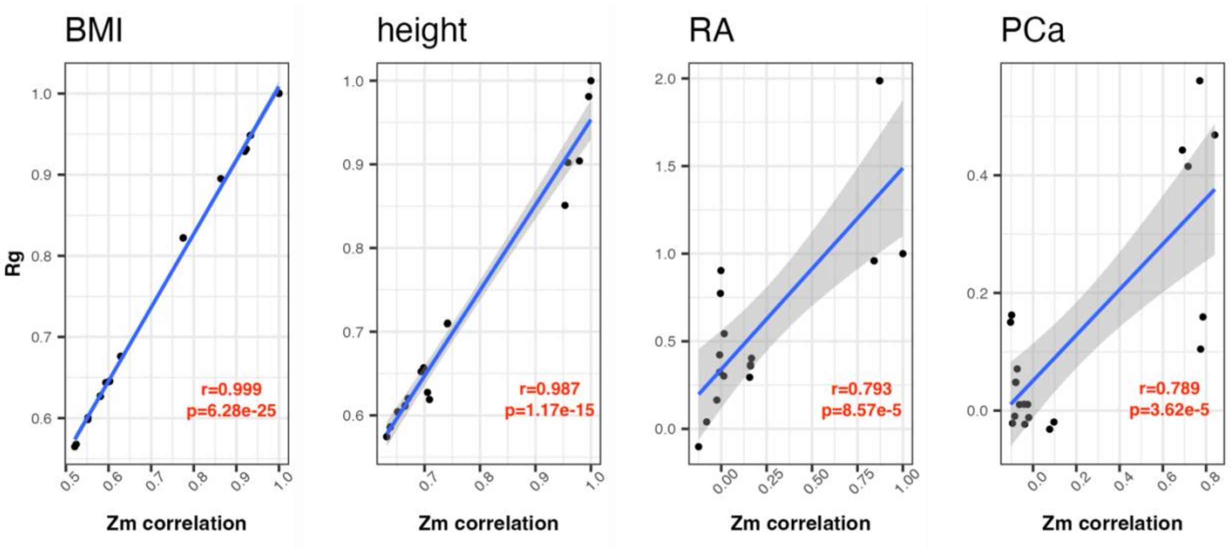


Figure S12. Genetic correlation between leading trait and focal trait is consistent with their factor scores correlation.

For each focal trait, we plotted its genetic correlation with 20 leading traits (displayed in **Figure 5**) in its leading factor (F1, F2, F86, F55) on y-axis. For this same set of leading trait – focal trait pair, we plotted their correlation in trait factor scores (standardized by posterior variances) across 100 factors on x-axis. As expected, factor scores correlation is concordant with genetic correlation. A regression line is fit to this relationship with grey confidence band.

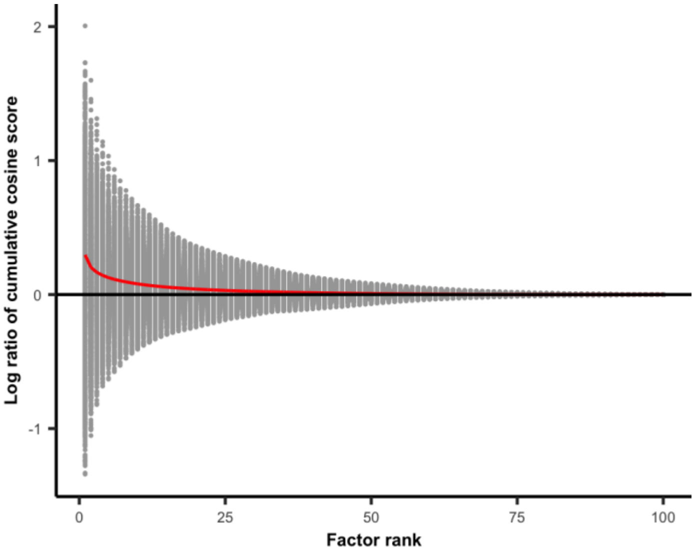


Figure S13. Cumulative squared cosine score for each trait was higher in FactorGo than in tSVD at each rank of pleiotropic factor.

We report the ratio of cumulative squared cosine score of FactorGo versus tSVD on log scale at each factor rank for each trait. The squared cosine score sums to 1 for each trait so that this ratio approaches log1=0 as rank increases. Red line is the mean of the ratio at each rank across traits.

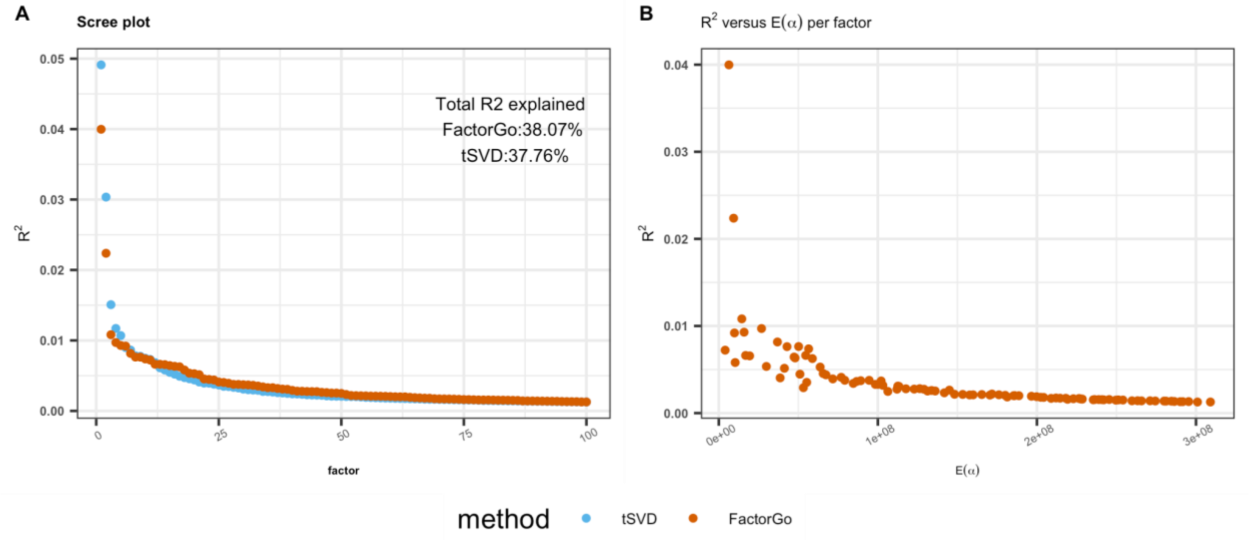


Figure S14. Variance explained by FactorGo factors tracks closely with posterior mean of ARD parameters α.

(A) Variance explained (R^2) by each factor from FactorGo and tSVD. The factors are ordered by R^2 for both methods. **(B)** Variance explained (R^2) by each factor in FactorGo versus their posterior mean of ARD prior parameter α.

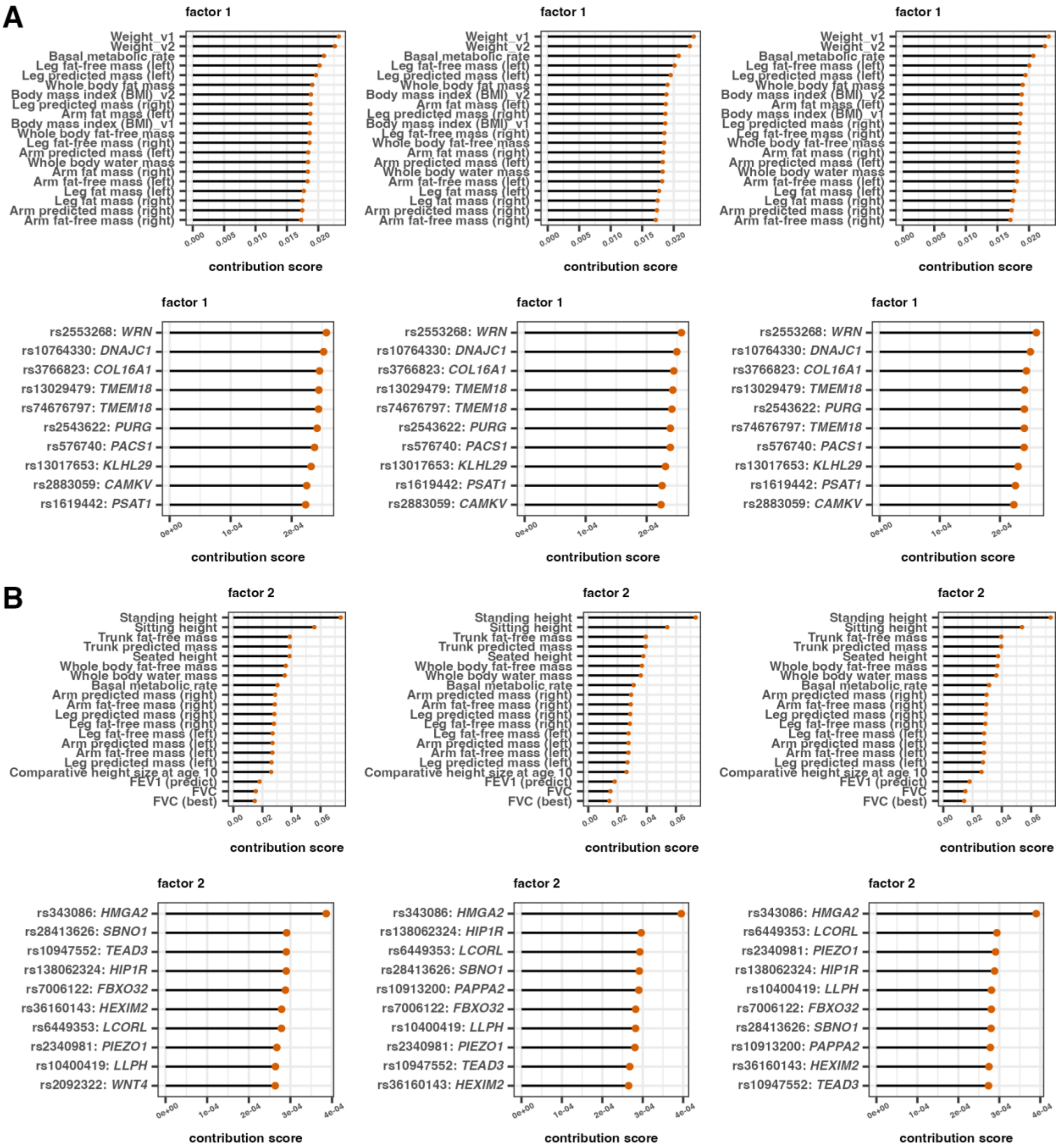


Figure S15. Top two leading factors in FactorGo are robust to choices of k.

A: Compare 20 leading traits and 10 leading variants in factor 1 across $k = 90{,}100{,}110$ from left to right.

B: Compare 20 leading traits and 10 leading variants in factor 2 across $k = 90{,}100{,}110$ from left to right.

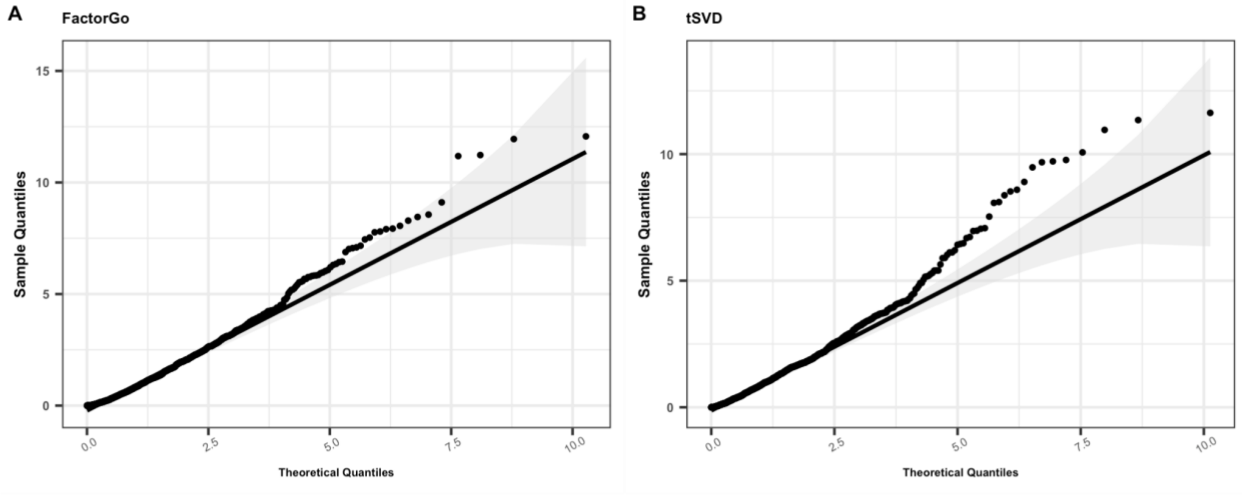


Figure S16. QQ plot of enrichment P values for randomly selected gene set annotations. To show our implementation of S-LDSC is well-calibrated for both **(A)** FactorGo and **(B)** tSVD, we compared P values distribution from S-LDSC enrichment results of 10 randomly selected gene sets to theoretical quantiles. Gray regions are pointwise confidence bands. Given the randomly selected gene sets of size ~2000 genes may be not truly “nuII”, there is some deviation from null distribution.

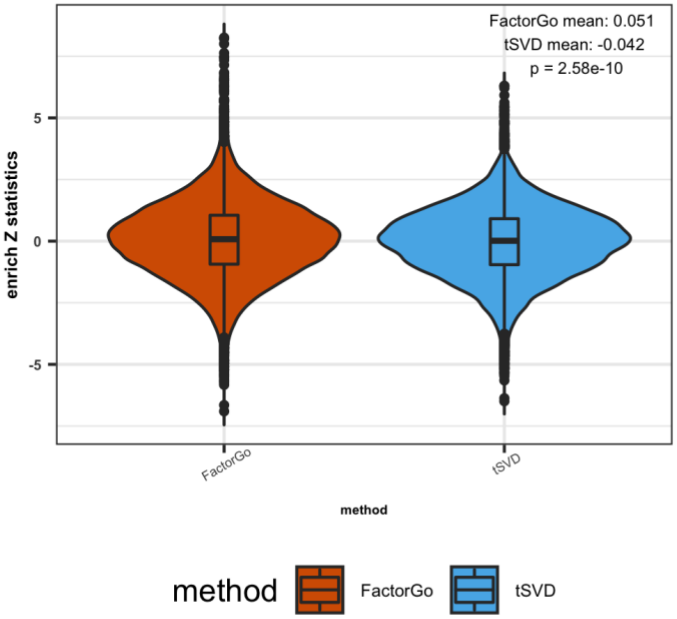


Figure S17. FactorGo factors show higher enrichment Z test statistics than tSVD. Violin plot and embedded boxplot of enrichment Z test statistics (one-sided test) from S-LDSC results for 205 annotations across all 100 factors for each method.

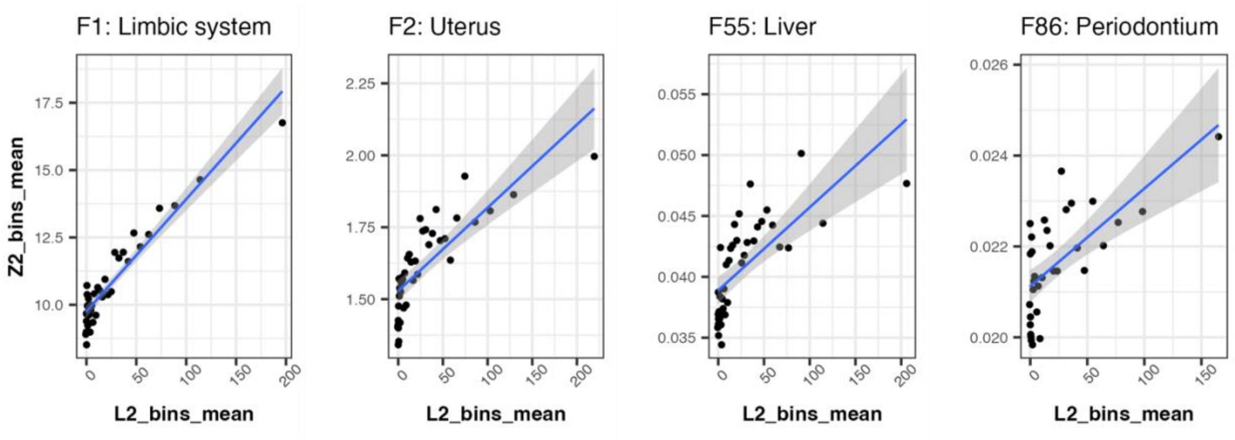


Figure S18. Pseudo-Z score has linear association with annotation LD scores.
For each pair of the leading factor and the leading enriched annotation for four focal traits, we calculated the mean squared pseudo-Z score in 50 binned annotation specific LD score bins.

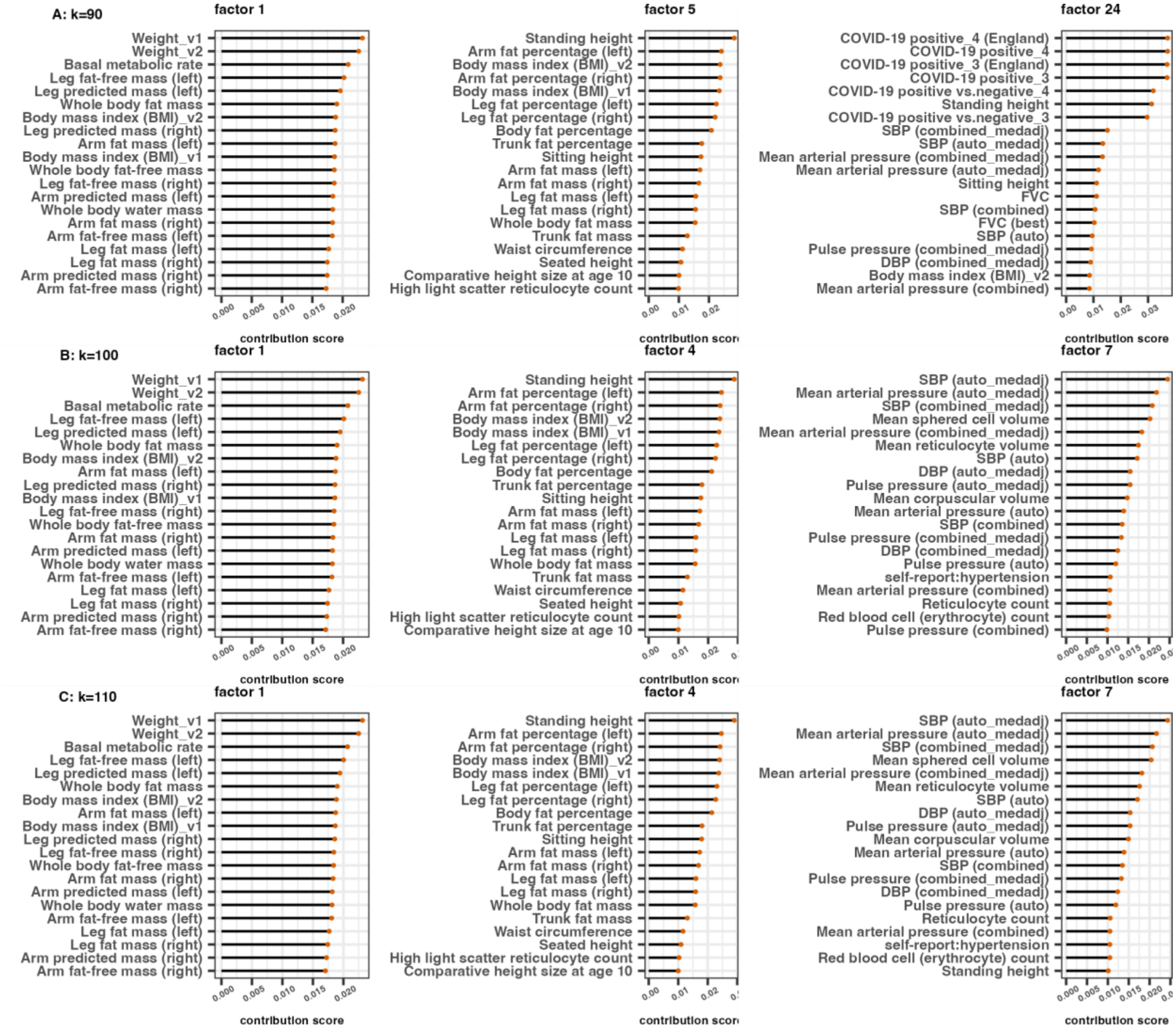


Figure S19. Three leading factors for BMI in FactorGo are robust to choices of k. Compare 20 leading traits in three leading factors in FactorGo results using (**A**) $k = 90$, (**B**) $k = 100$, (**C**) $k = 110$.

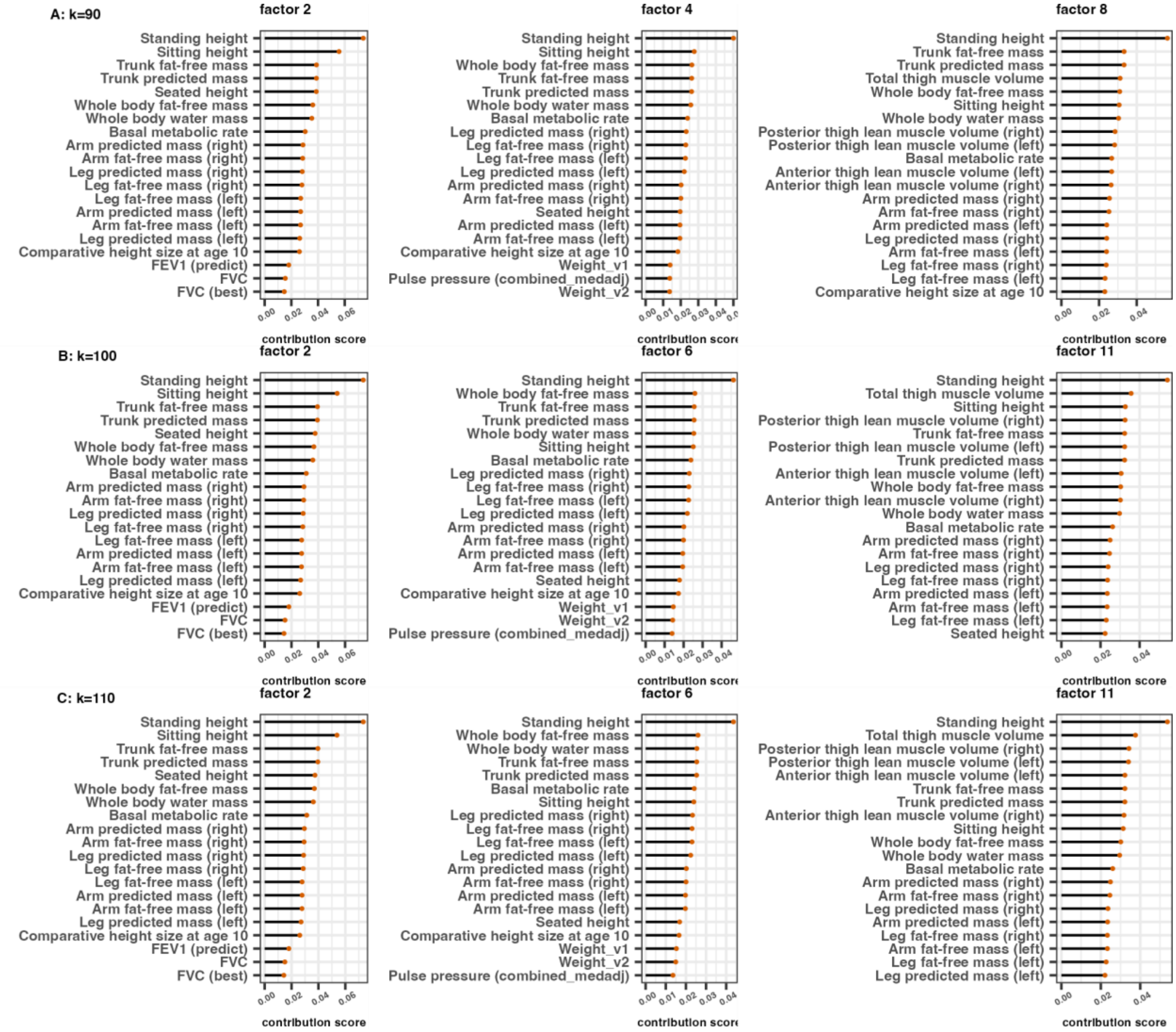


Figure S20. Three leading factors for height in FactorGo are robust to choices of k. Compare 20 leading traits in three leading factors in FactorGo results using (**A**) $k = 90$, (**B**) $k = 100$, (**C**) $k = 110$.

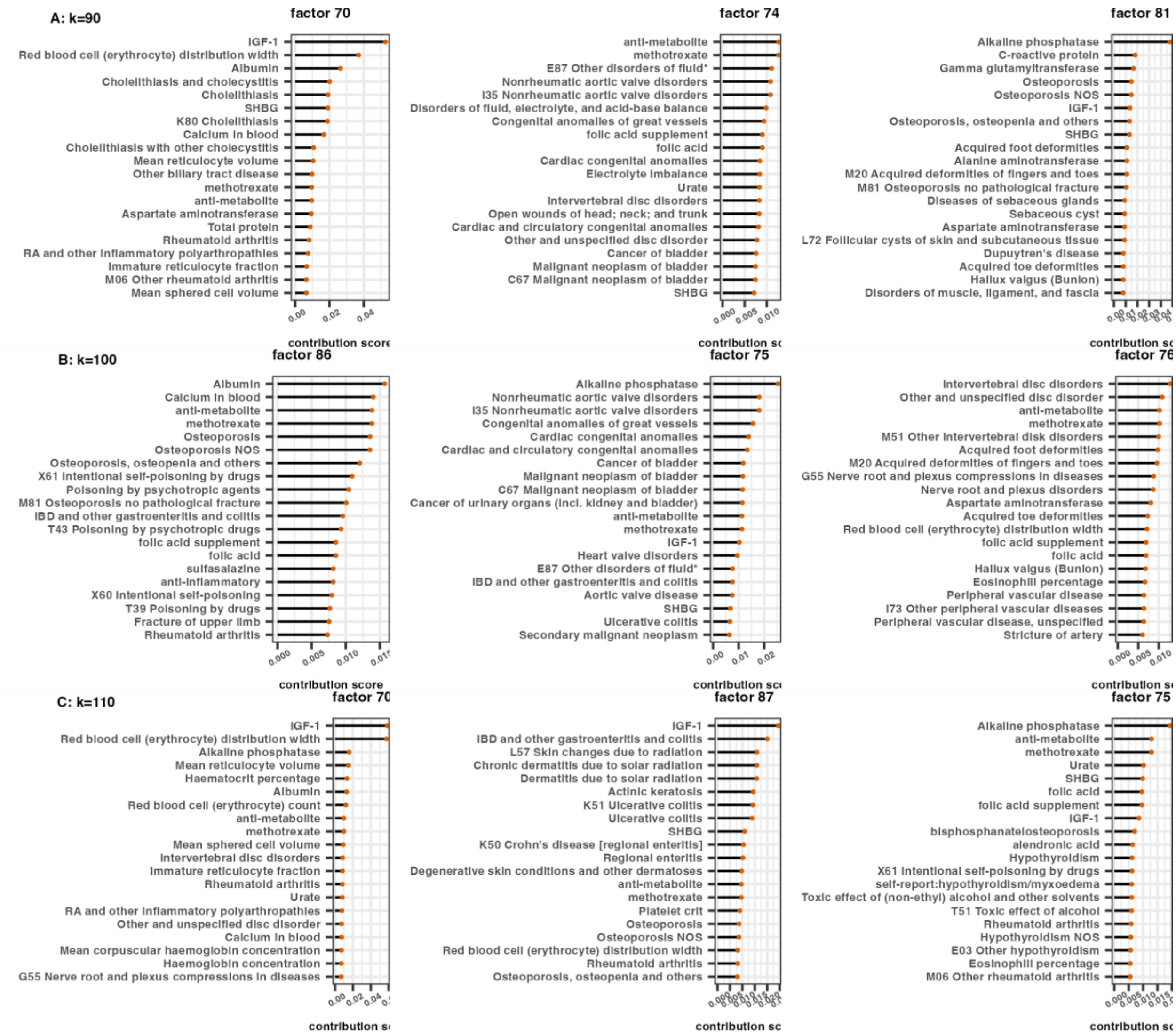


Figure S21. Three leading factors for RA in FactorGo are overall robust to choices of k. Compare 20 leading traits in three leading factors in FactorGo results using (**A**) $k = 90$, (**B**) $k = 100$, (**C**) $k = 110$.

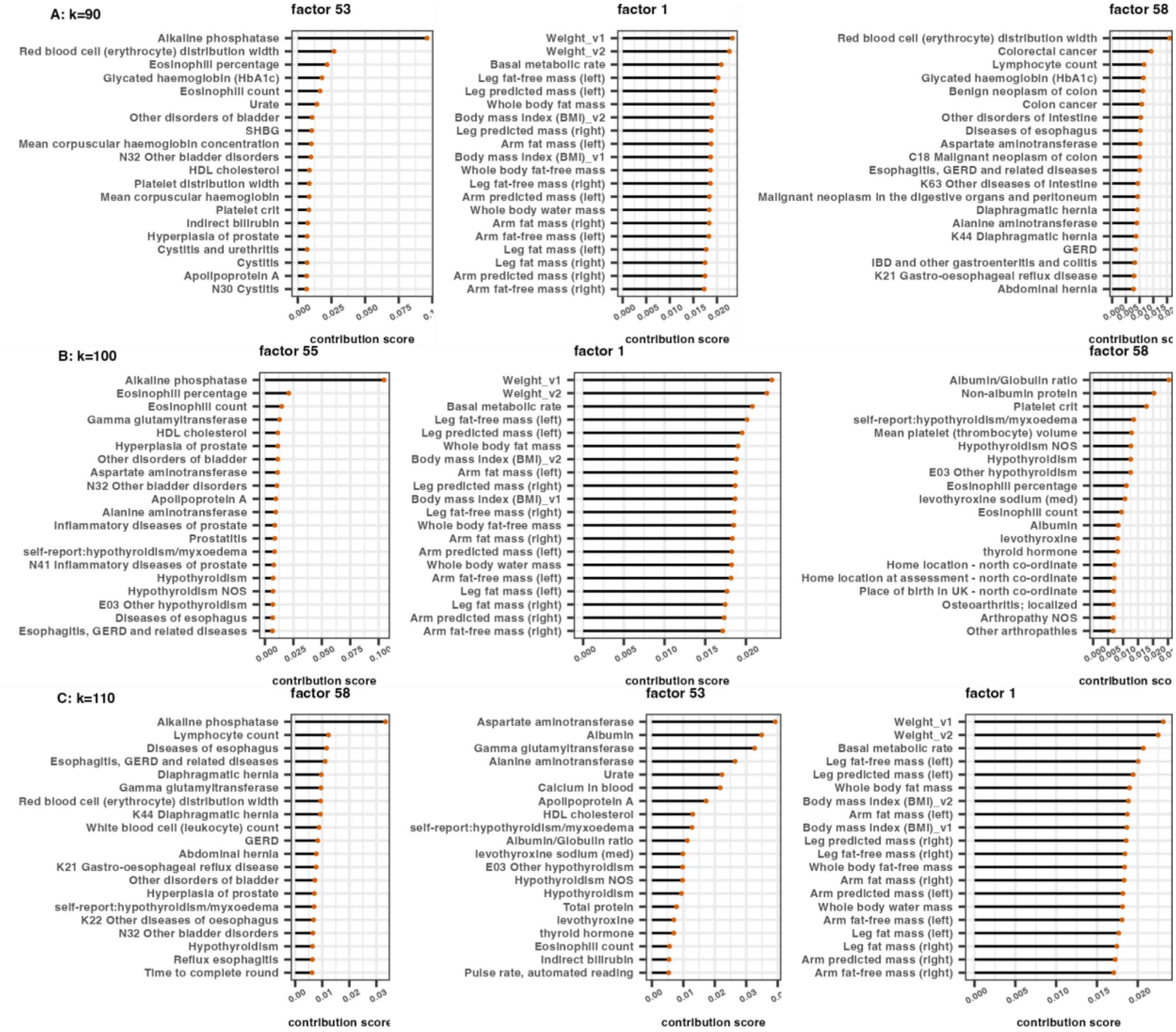


Figure S22. Three leading factors for PCa in FactorGo are overall robust to choices of k. Compare 20 leading traits in three leading factors in FactorGo results using (**A**) $k = 90$, (**B**) $k = 100$, (**C**) $k = 110$.

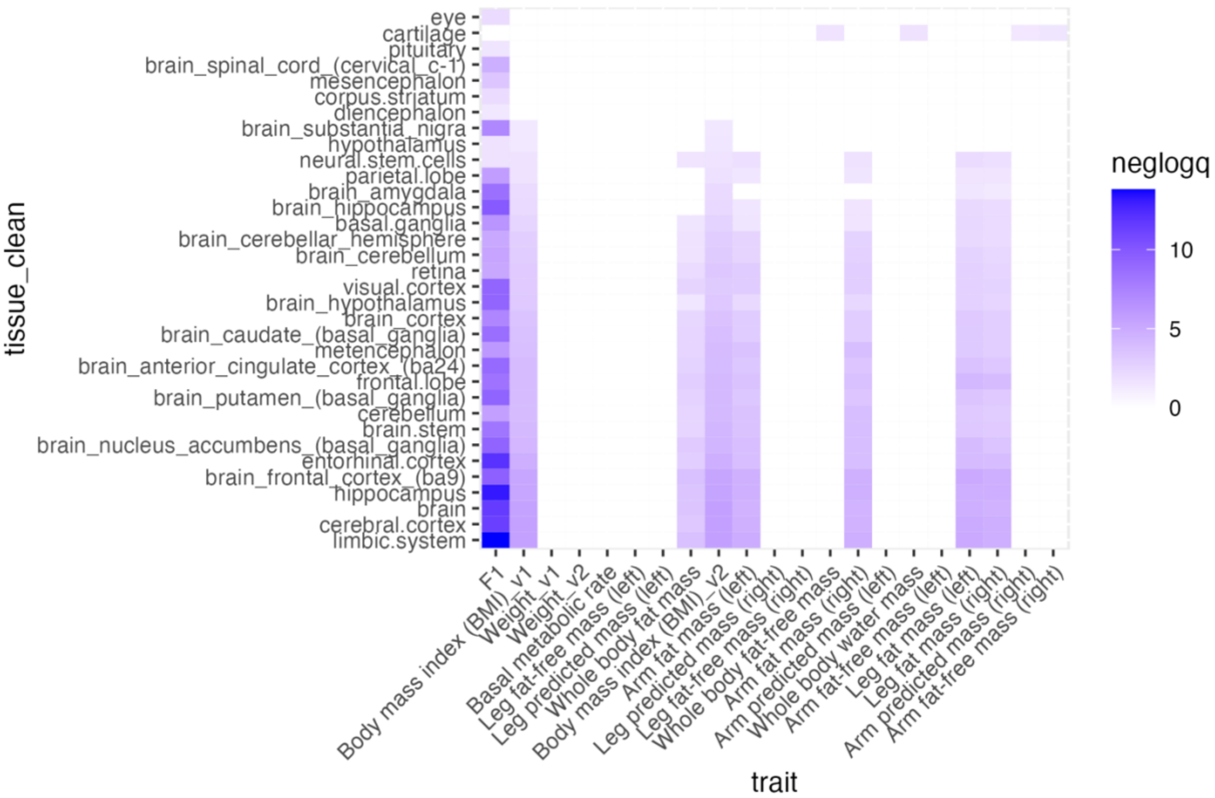


Figure S23. The leading factor for BMI shows consistent brain tissue enrichment with leading trait enrichments.

We plotted $-\log_{10}(qvalue)$ of enriched tissue at FDR < 0.05 for the leading factor (F1), BMI trait, and 20 leading traits.

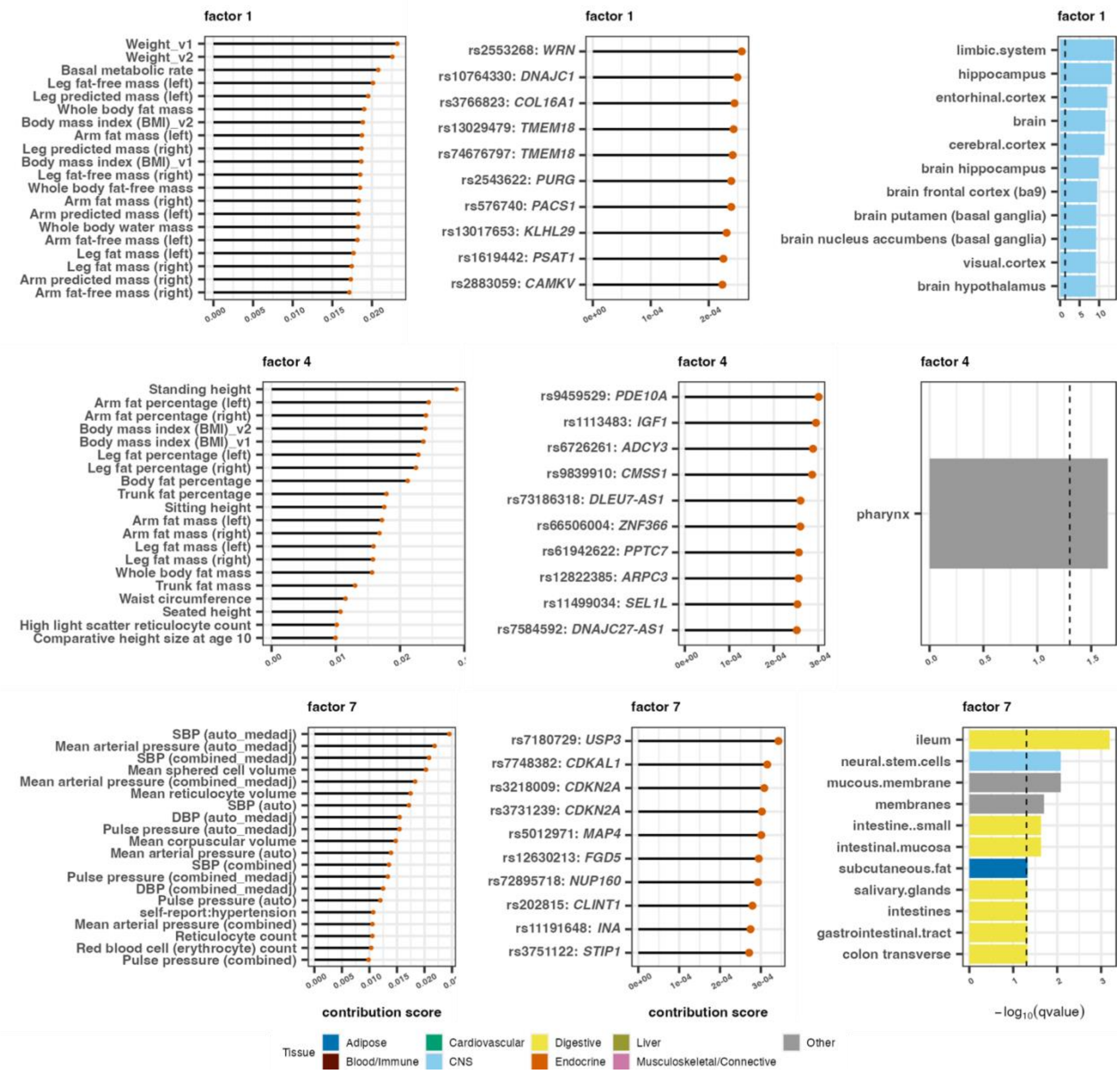


Figure S24. Characterizing three leading factors in FactorGo for BMI.

Results for factor 1, 4 and 7 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type (truncated to 10 if more than 20 enriched annotations). Dashed lines are FDR threshold at 0.05. Detailed results in **Table S4**.

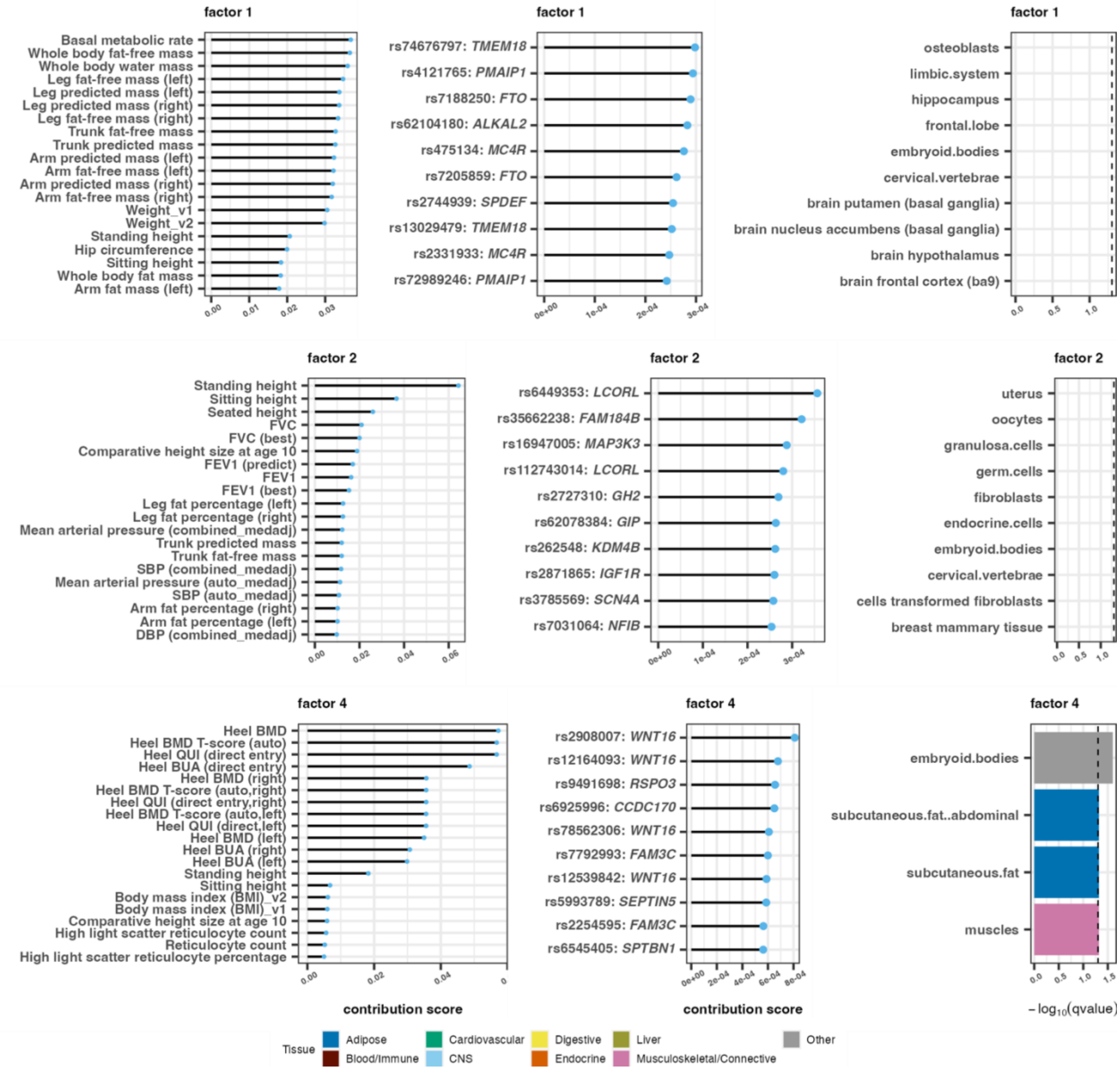


Figure S25. Characterizing three leading factors in tSVD for BMI.
Results for factor 1, 2 and 4 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed result in **Table S5**.

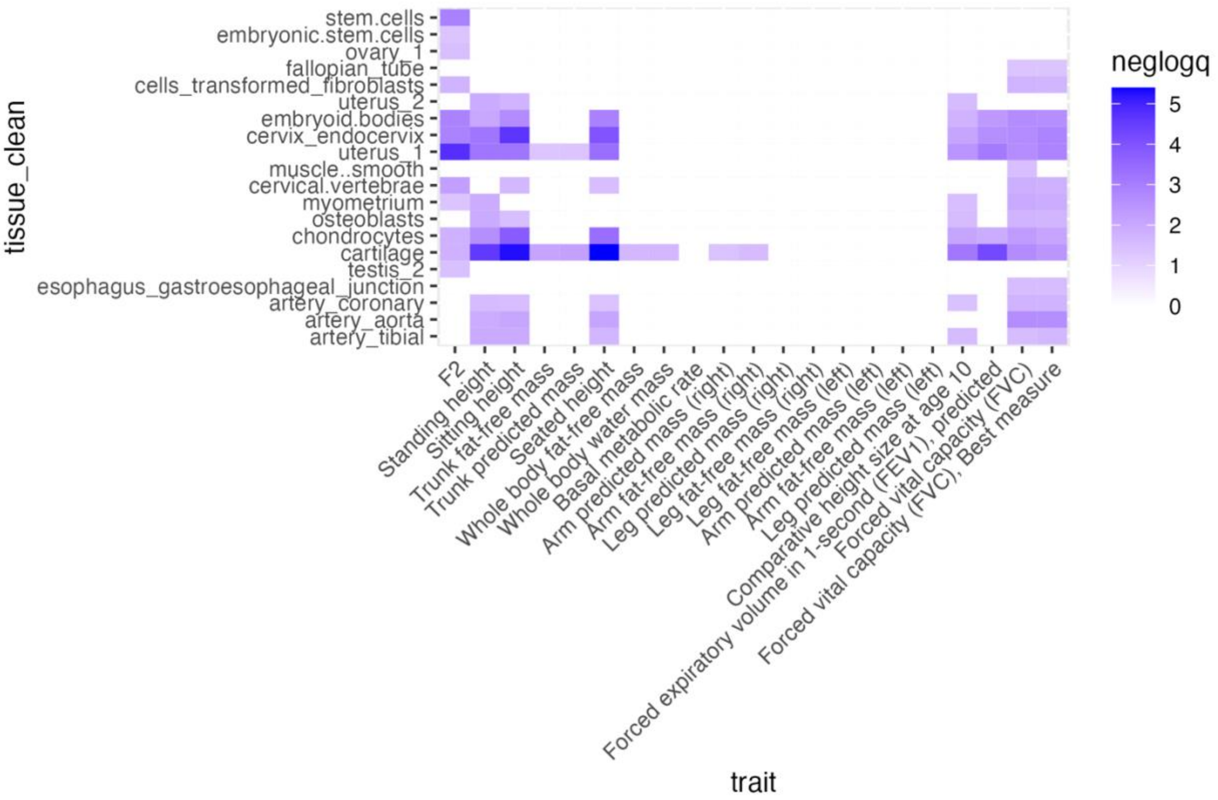


Figure S26. The leading factor for standing height shows consistent tissue enrichment with single trait enrichment.

We plotted $-\log_{10}(qvalue)$ of enriched tissue at FDR < 0.05 for the leading factor (F2), height, and 20 leading traits.

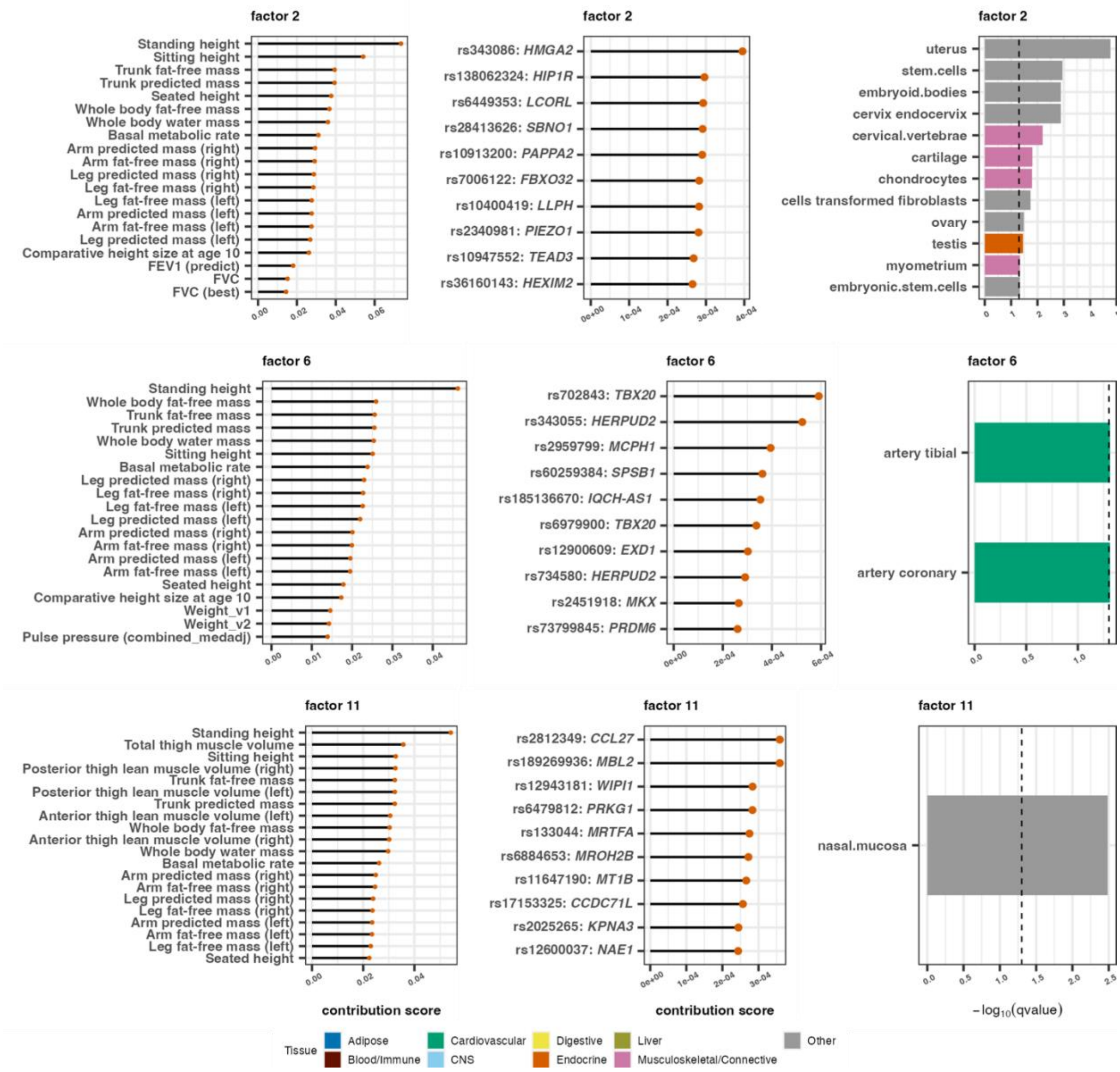


Figure S27. Characterizing three leading factors in FactorGo for height.
Results for factor 2, 6 and 11 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed results in **Table S4**.

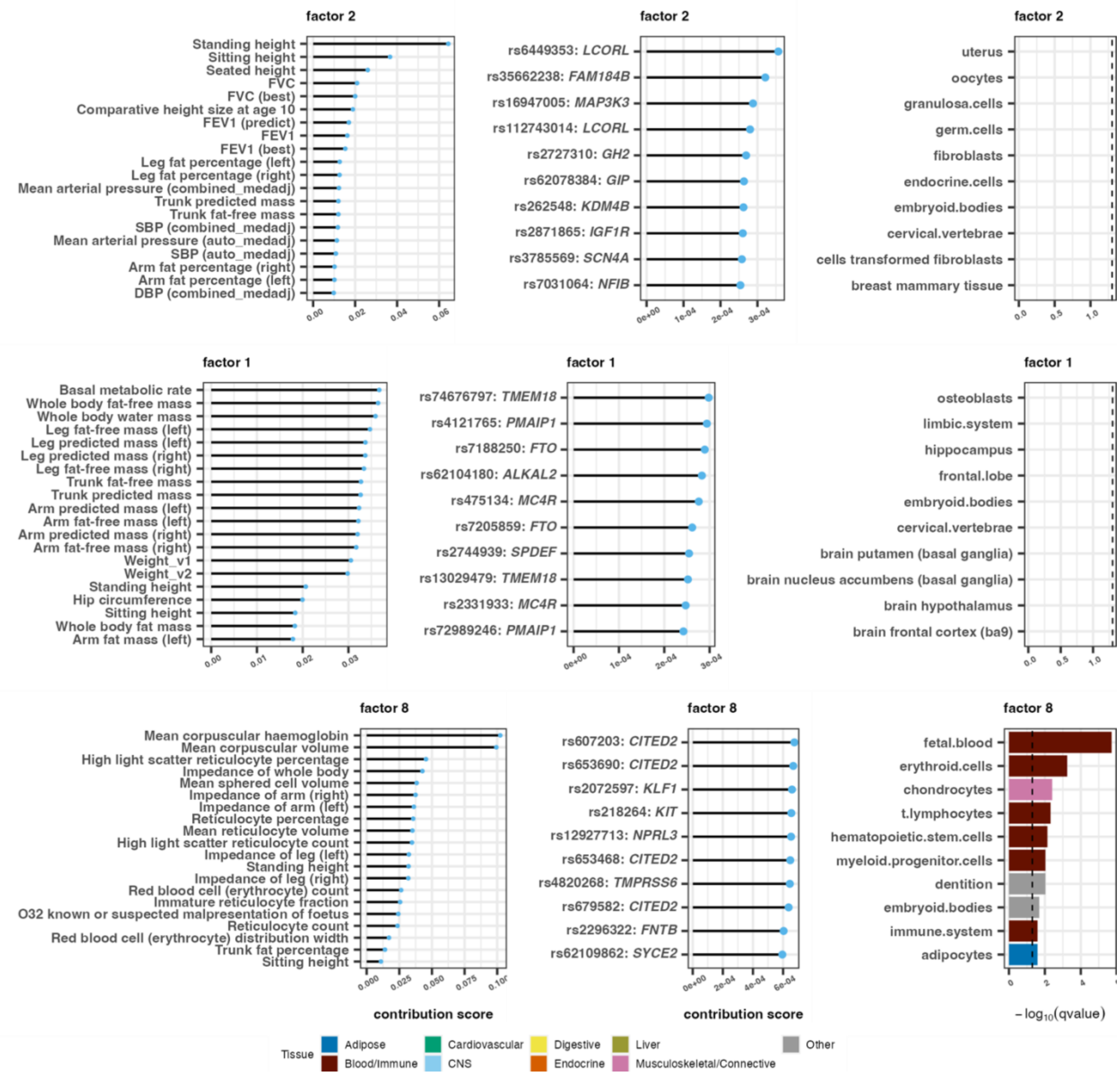


Figure S28. Characterizing three leading factors in tSVD for height.
Results for factor 2, 1 and 8 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed results in **Table S5**.

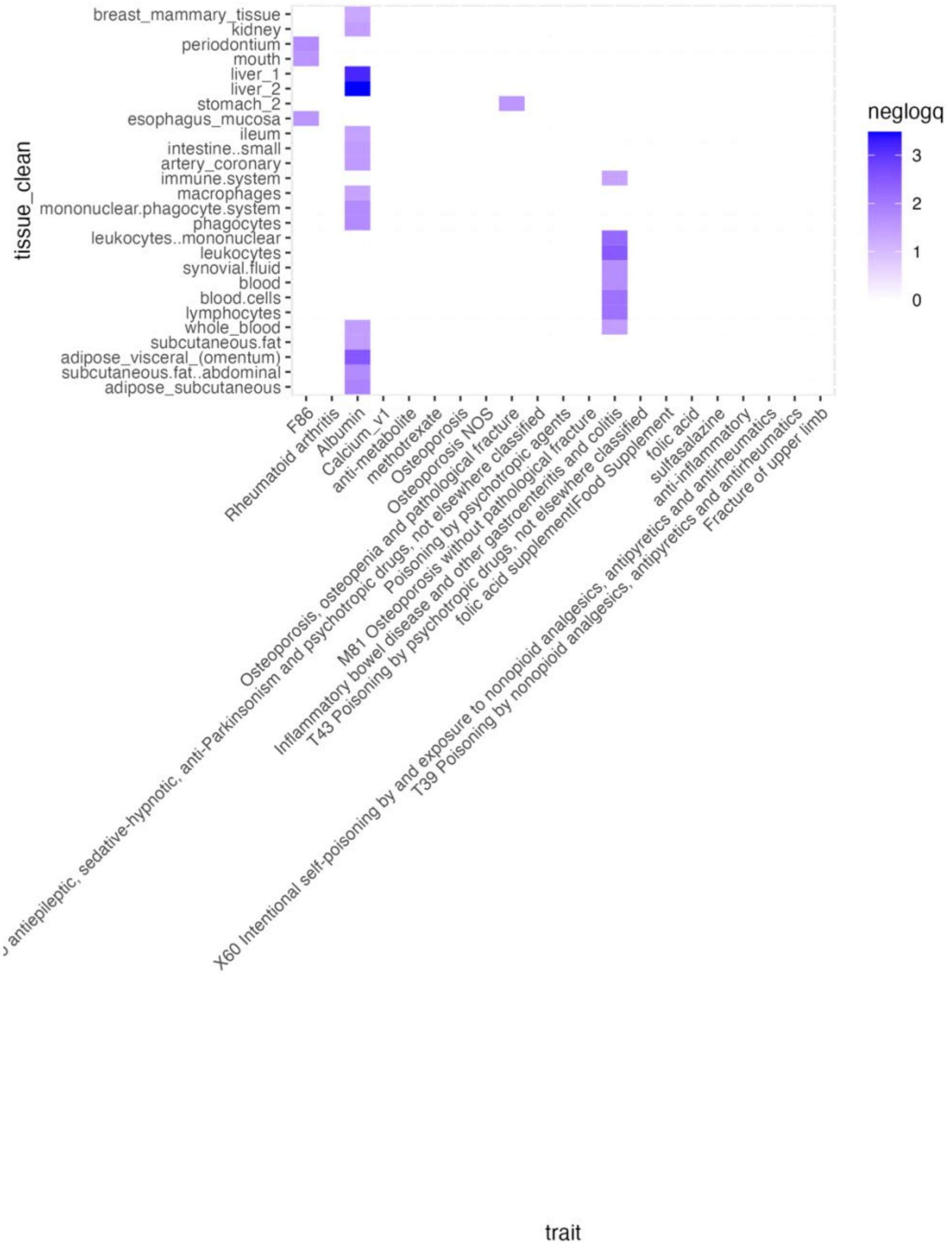


Figure S29. The leading factor for RA identifies novel shared tissue enrichment that is not found using single trait genome-wide variants.

We plotted $-\log_{10}(qvalue)$ of enriched tissue at FDR < 0.05 for the leading factor (F86), RA and 20 leading traits.

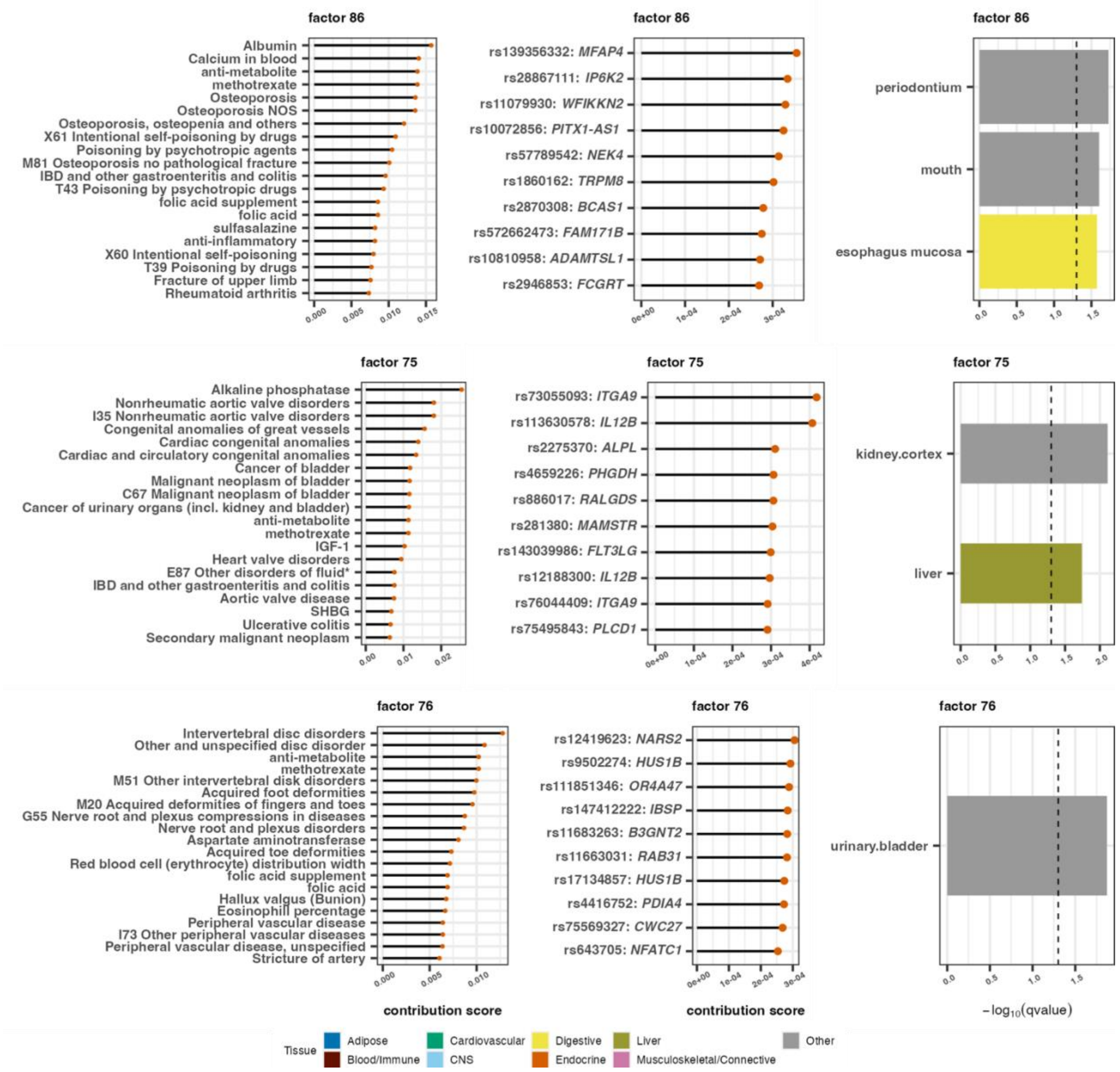


Figure S30. Characterizing three leading factors in FactorGo for RA.
Results for factor 86, 75 and 76 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed results in **Table S4**.

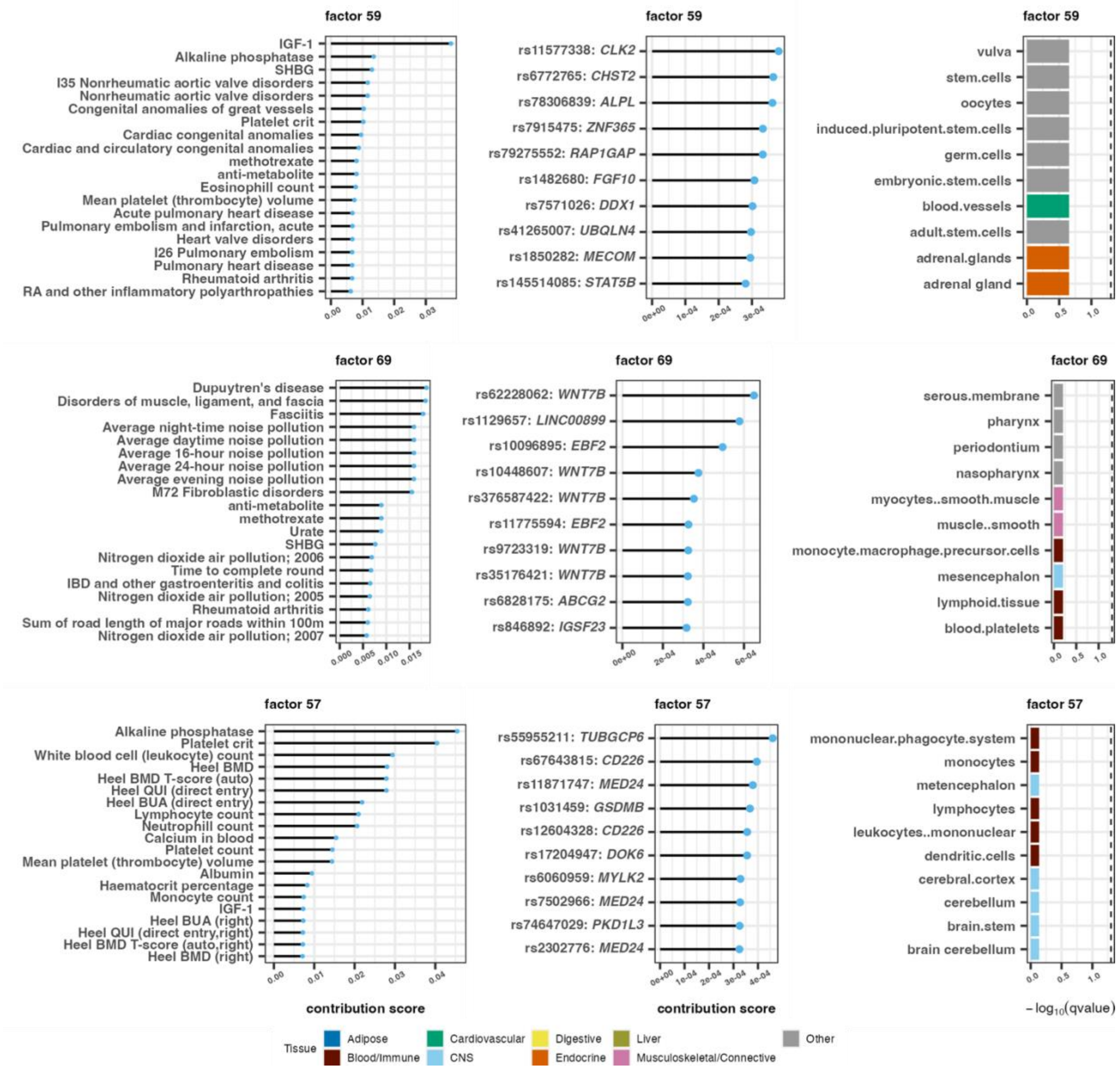


Figure S31. Characterizing three leading factors in tSVD for RA.

Results for factor 59, 69 and 57 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed results in **Table S5**.

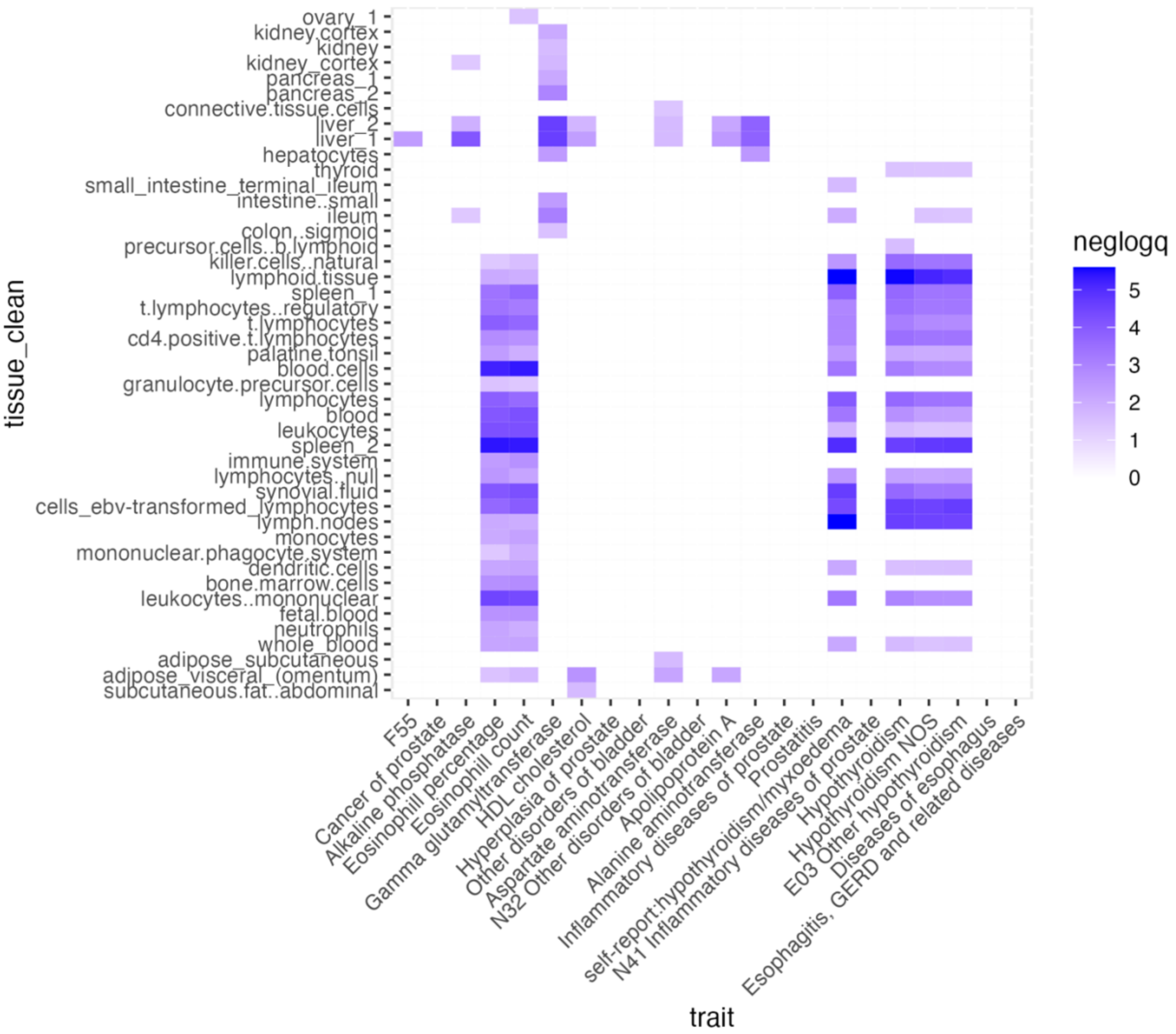


Figure S32. Leading factor for PCa shows consistent liver enrichment with single trait results.

We plotted $-\log_{10}(qvalue)$ of enriched tissue at FDR < 0.05 for the leading factor (F55), PCa and 20 leading traits.

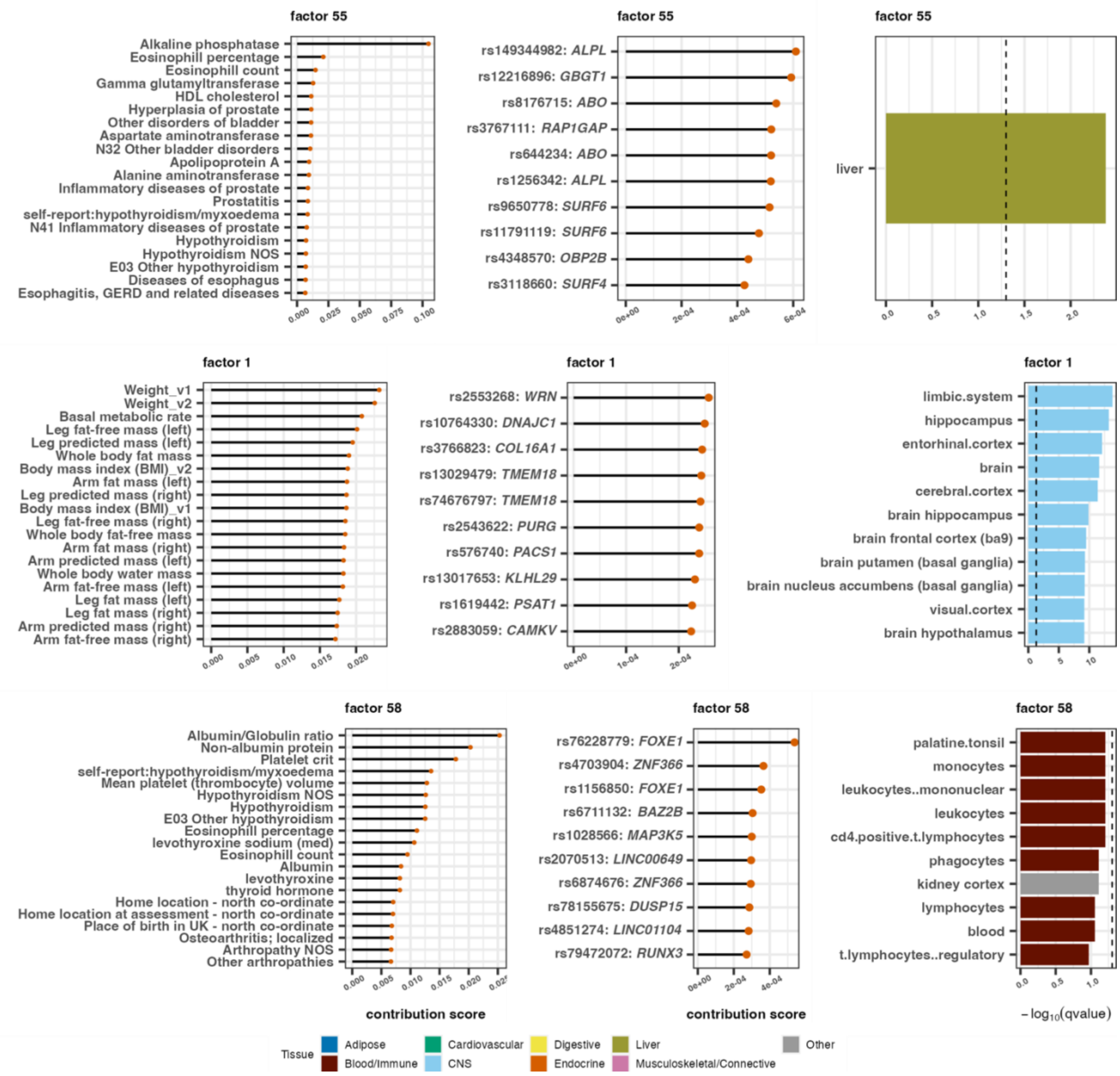


Figure S33. Characterizing three leading factors in FactorGo for PCa.
Results for factor 55, 1 and 58 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed result in **Table S4**.

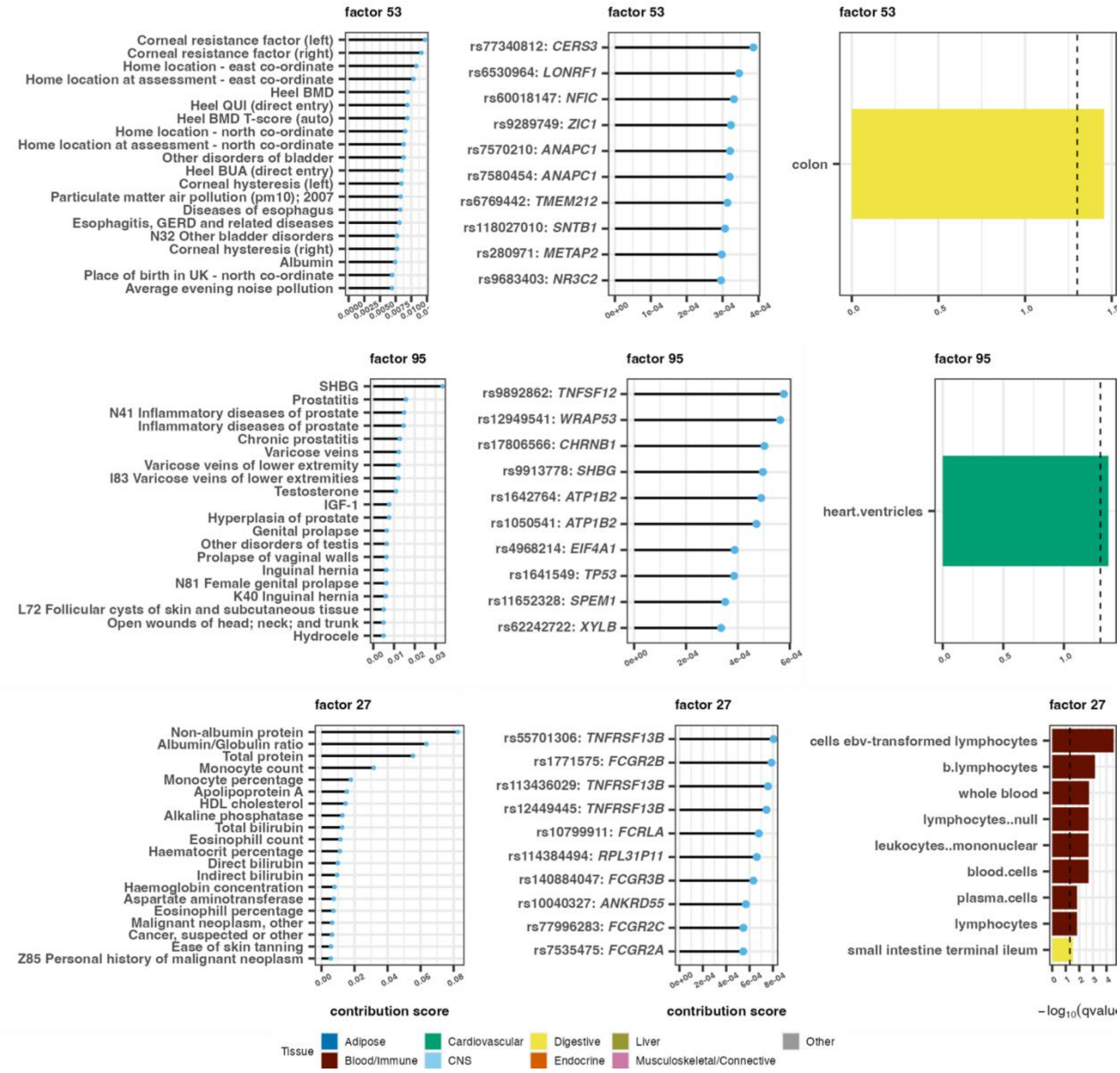


Figure S34. Characterizing three leading factors in tSVD for PCa.

Results for factor 53, 95 and 27 (row) include 20 leading traits, 10 leading variants with closest gene, and enriched LDSC-SEG tissue or cell type. Dashed lines are FDR threshold at 0.05. Detailed results in **Table S5**.

Supplemental Tables

Group	Type	Abbreviation	Number of traits	Example	UKB Field ID
Disease	BIN	Disease (D)	1130	Type 2 diabetes	ICD10, phecode, self-reported 20002, COVID19
Cancer	BIN	Cancer	20	Breast Cancer	self-reported 20001, phecode
Family history	BIN	Family history (FH)	26	Illness of mother, illness of father	20107, 20110
Treatment/Medication/Prescription	BIN	Medication/Prescription (MED)	501	Aspirin	20003, 42039 harmonized
Physical measures	QT	Physical measures (PM)	202	Weight, Pulse rate	Category 100003, 100006, derived variables
Mental health	QT	Mental Health (MENT)	94	General happiness, prospective memory	Category 100026, 100059, 136
Biological samples (eg. assay)	QT	Assay	67	Cholesterol, Monocyte count	Category 100078
Questionnaire (eg.food intake, exercise, environment)	QT	Lifestyle and Exposure (LIFE)	418	Milk intake, Smoking, Physical activity	Category 100090, 100025, 113,123
Misc	QT	Miscellaneous (MISC)	25	Birth weight, number of operations, home location	

Table S1. Groups of 2,483 phenotypes

Phenocode	Description
22410	Total trunk fat volume
23099	Body fat percentage
23100	Whole body fat mass
23111	Leg fat percentage (right)
23112	Leg fat mass (right)
23115	Leg fat percentage (left)
23116	Leg fat mass (left)
23119	Arm fat percentage (right)
23120	Arm fat mass (right)
23123	Arm fat percentage (left)
23124	Arm fat mass (left)
23127	Trunk fat percentage
23128	Trunk fat mass

Table S2. 13 Body fat mass traits.
Phenocode is a field ID described by UKB.

Phenocode	Description
M81	M81 Osteoporosis without pathological fracture
743	Osteoporosis, osteopenia and pathological fracture
743.1	Osteoporosis
743.11	Osteoporosis NOS
20002	self-report:osteoporosis

Table S3. 5 Osteoporosis traits
Phenocode is a field ID described by UKB. NOS: not otherwise specified.

Supplemental Reference

1. Yong, A.G., and Pearce, S. (2013). A beginner’s guide to factor analysis: Focusing on exploratory factor analysis. Tutor. Quant. Methods Psychol. *9*, 79–94.

2. Bishop, C.M. (1999). Variational principal components. 509–514.

3. Hansen, P.C. (1987). The truncatedSVD as a method for regularization. BIT *27*, 534–553.

4. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. *112*, 859–877.

5. Luttinen, J., and Ilin, A. (2010). Transformations in variational Bayesian factor analysis to speed up learning. Neurocomputing *73*, 1093–1102.

6. Nawrocki, A.R., Rodriguez, C.G., Toolan, D.M., Price, O., Henry, M., Forrest, G., Szeto, D., Keohane, C.A., Pan, Y., Smith, K.M., et al. (2014). Genetic deletion and pharmacological inhibition of phosphodiesterase 10A protects mice from diet-induced obesity and insulin resistance. Diabetes *63*, 300–311.

7. Berryman, D.E., Glad, C.A.M., List, E.O., and Johannsson, G. (2013). The GH/IGF-1 axis in obesity: pathophysiology and therapeutic considerations. Nat. Rev. Endocrinol. *9*, 346–356.

8. Busetto, L., Calo’, E., Mazza, M., De Stefano, F., Costa, G., Negrin, V., and Enzi, G. (2009). Upper airway size is related to obesity and body fat distribution in women. Eur. Arch. Otorhinolaryngol. *266*, 559–563.

9. Landi, F., Calvani, R., Picca, A., Tosato, M., Martone, A.M., Ortolani, E., Sisto, A., D’Angelo, E., Serafini, E., Desideri, G., et al. (2018). Body Mass Index is Strongly Associated with Hypertension: Results from the Longevity Check-up 7+ Study. Nutrients *10*,.

10. Liu, P.-H., Wu, K., Ng, K., Zauber, A.G., Nguyen, L.H., Song, M., He, X., Fuchs, C.S., Ogino, S., Willett, W.C., et al. (2019). Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women. JAMA Oncol *5*, 37–44.

11. Kirk, E.P., Sunde, M., Costa, M.W., Rankin, S.A., Wolstein, O., Castro, M.L., Butler, T.L., Hyun, C., Guo, G., Otway, R., et al. (2007). Mutations in Cardiac T-Box Factor Gene TBX20 Are Associated with Diverse Cardiac Pathologies, Including Defects of Septation and Valvulogenesis and Cardiomyopathy. Am. J. Hum. Genet. *81*, 280–291.

12. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

13. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J. ’an, Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

14. Davila, M.L., Xu, M., Huang, C., Gaddes, E.R., Winter, L., Cantorna, M.T., Wang, Y., and Xiong, N. (2022). CCL27 is a crucial regulator of immune homeostasis of the skin and mucosal tissues. iScience *25*, 104426.

15. Degn, S.E., Jensenius, J.C., and Thiel, S. (2011). Disease-causing mutations in genes of

the complement system. Am. J. Hum. Genet. *88*, 689–705.

16. Urlacher, S.S., Ellison, P.T., Sugiyama, L.S., Pontzer, H., Eick, G., Liebert, M.A., Cepon-Robins, T.J., Gildner, T.E., and Snodgrass, J.J. (2018). Tradeoffs between immune function and childhood growth among Amazonian forager-horticulturalists. Proc. Natl. Acad. Sci. U. S. A. *115*, E3914–E3921.

17. Abraham, S., Begum, S., and Isenberg, D. (2004). Hepatic manifestations of autoimmune rheumatic diseases. Ann. Rheum. Dis. *63*, 123–129.

18. Anders, H.-J., and Vielhauer, V. (2011). Renal co-morbidity in patients with rheumatic diseases. Arthritis Research & Therapy *13*, 222.

19. Wilson, R.L., Taaffe, D.R., Newton, R.U., Hart, N.H., Lyons-Wall, P., and Galvão, D.A. (2022). Obesity and prostate cancer: A narrative review. Crit. Rev. Oncol. Hematol. *169*, 103543.

20. Deng, T., Lyon, C.J., Bergin, S., Caligiuri, M.A., and Hsueh, W.A. (2016). Obesity, Inflammation, and Cancer. Annu. Rev. Pathol. *11*, 421–449.

21. Fernández, L.P., López-Márquez, A., Martínez, A.M., Gómez-López, G., and Santisteban, P. (2013). New insights into FoxE1 functions: identification of direct FoxE1 targets in thyroid cells. PLoS One *8*, e62849.