iScience, Volume *26*

Supplemental information

SuSiE PCA: A scalable Bayesian variable selection technique for principal component analysis

Dong Yuan and Nicholas Mancuso

1 Supplementary Figures

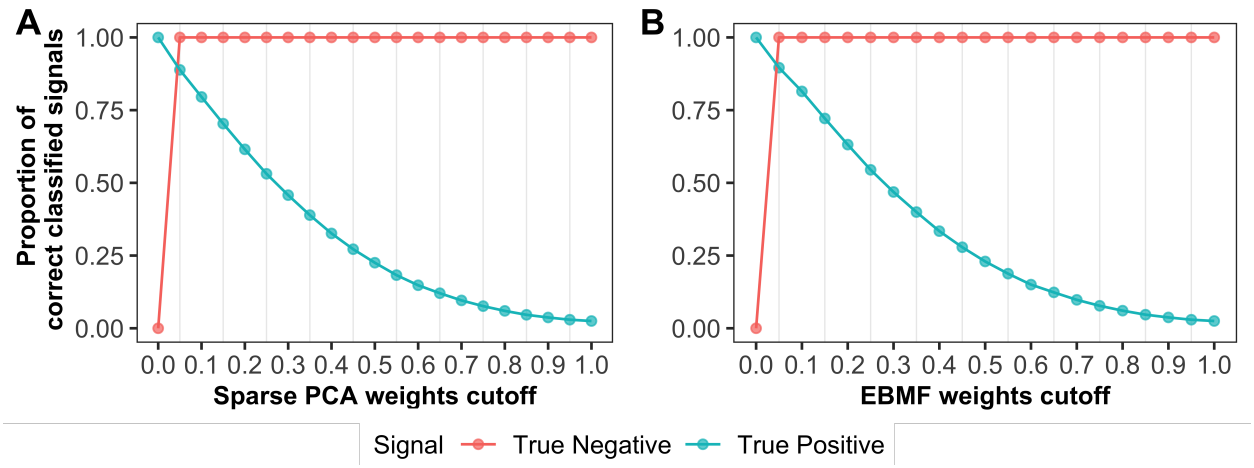

Figure S1. Sensitivity decreases fast as cutoff value increases when choosing weights from sparse PCA and EBMF for variable selection, related to Figure 1
The proportion of correct classified signals using posterior weights from sparse PCA (A) and EBMF (B) as the cutoff. The green dots represent sensitivity, i.e. $Pr(\text{weights} \geq \text{cutoff} \,|\, \text{True positive signal})$, the red dots represent specificity, i.e. $Pr(\text{weights} < \text{cutoff} \,|\, \text{True false signal})$. For consistency and comparable between PIPs and weights, the weights are standardized to be ranged from 0 to 1.

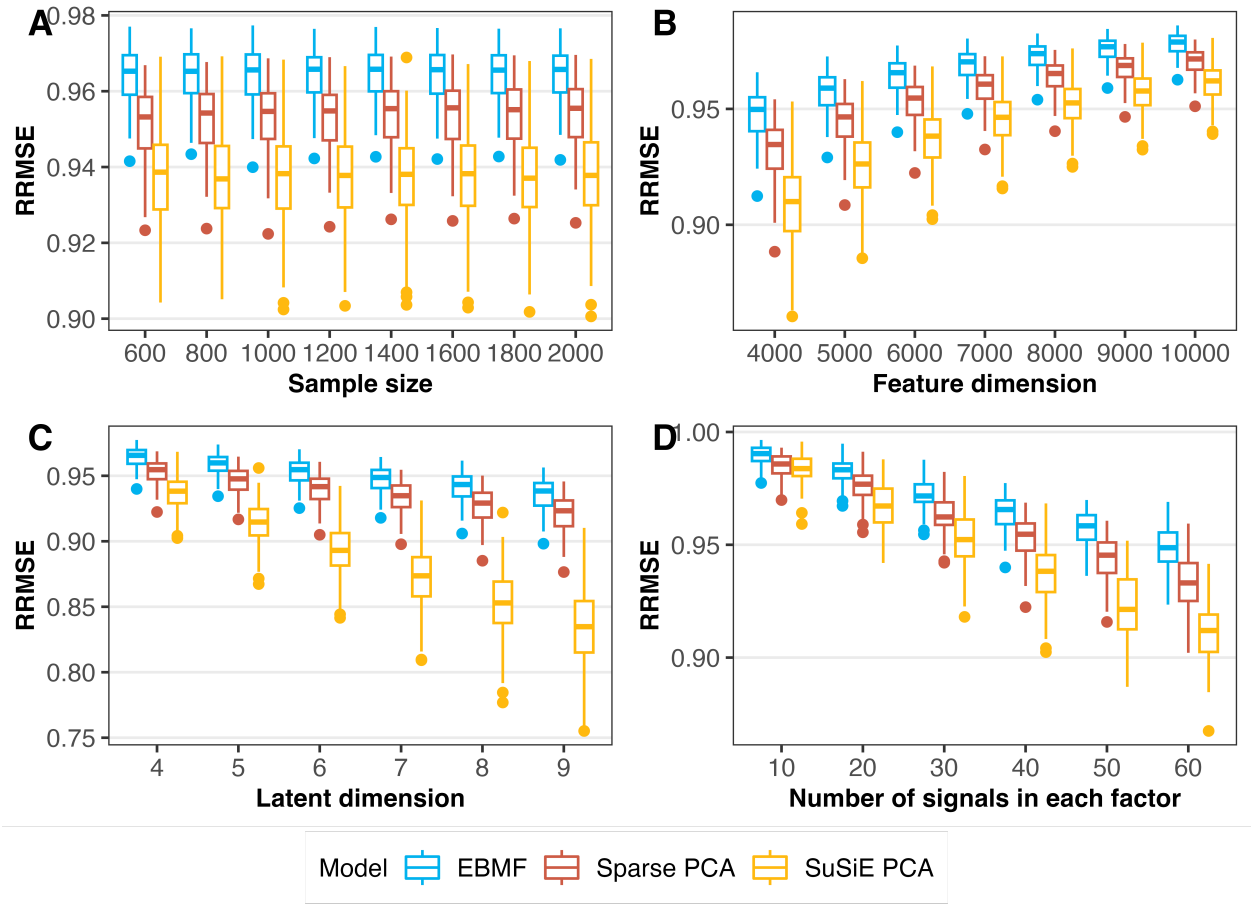

Figure S2. SuSiE PCA has the lowest RRMSE across all simulation settings, related to Figure 2.

The RRMSE is defined in **Simulation** equation (2.31), which is an assessment of the model prediction performance. The base simulation data is the same as the simulation setting in Figure 1. For each scenario in (A-D) we only vary one of the parameters at a time to generate the simulation data while fixing the other 3 parameters and then input the true parameters (N,P,K,L) into models. Finally, we compute the RRMSE based on equation (2.31) and plot them as a function of N, P, K, L.

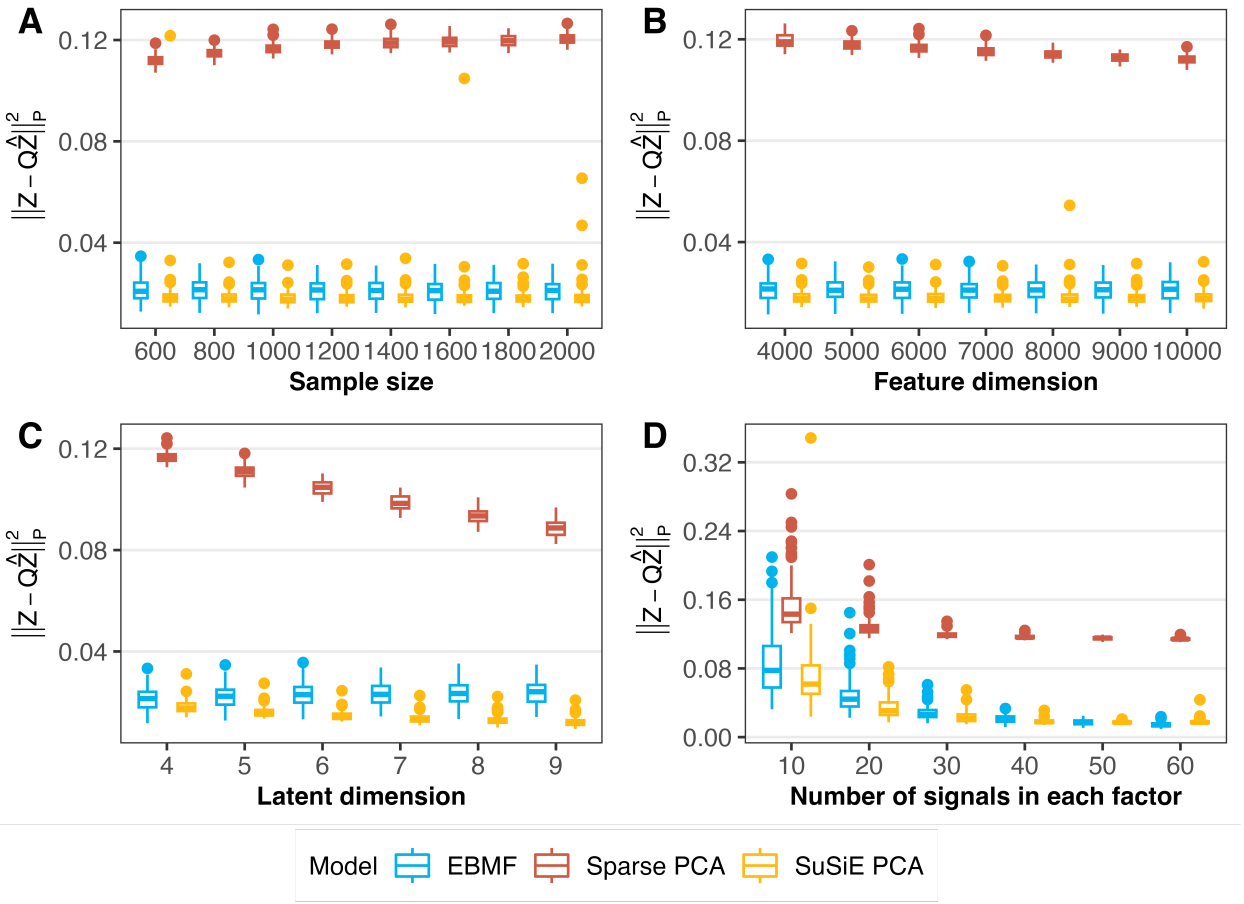

Figure S3. SuSiE PCA and EBMF has lower Procrustes error of latent factor Z than sparse PCA across all simulation settings, related to Figure 2.
The Procrustes error for latent factor Z is computed in the same manner of loading (Figure 1) using equation (2.30). For each scenario in (A-D) we only vary one of the parameters at a time to generate the simulation data while fixing the other 3 parameters and then input the true parameters (N,P,K,L) into models. Finally, we compute the Procrustes errors of $\mathbf{Z}$ based on equation (2.30) and plot them as a function of N, P, K, L.

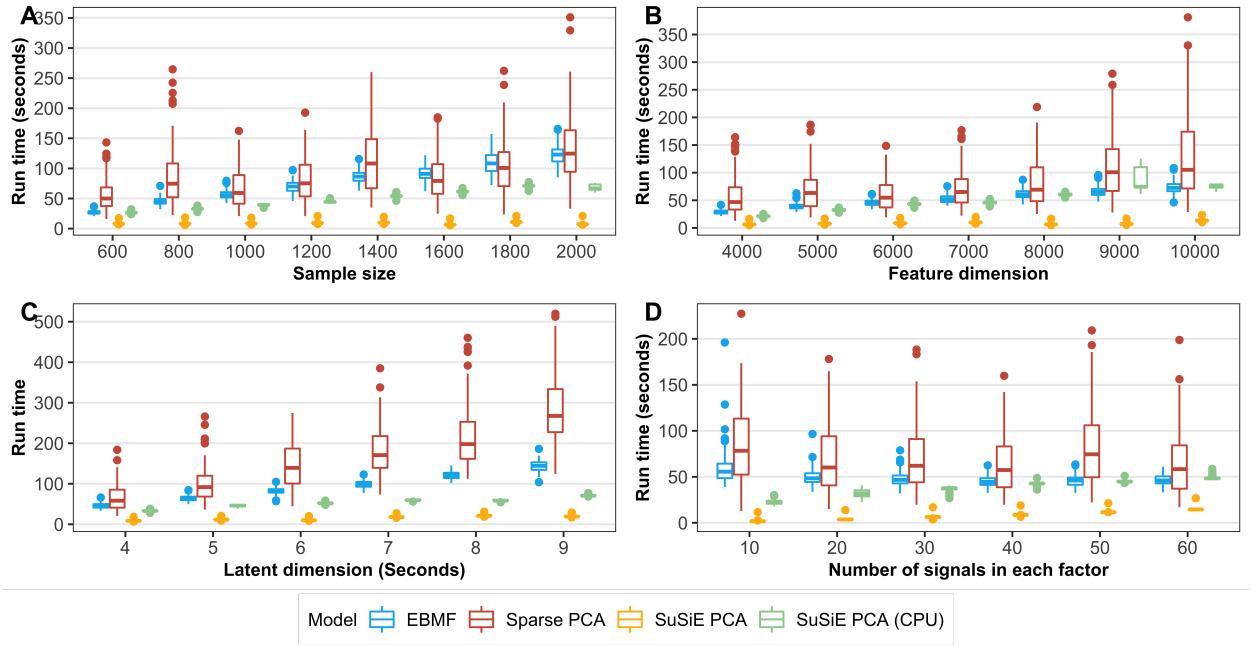

Figure S4. SuSiE PCA remains the fastest method on either CPU or GPU than sparse PCA and EBMF across all simulation settings, related to Figure 2 and Table 1.
All analyses are performed on high-performance computing center with the same CPU (AMD EPYC 7302 16-Core Processor) or GPU (Nvidia Tesla A40). Noticed that platform where we collect the runtime data in this figure is different from that in **Table 1** and therefore is not comparable.

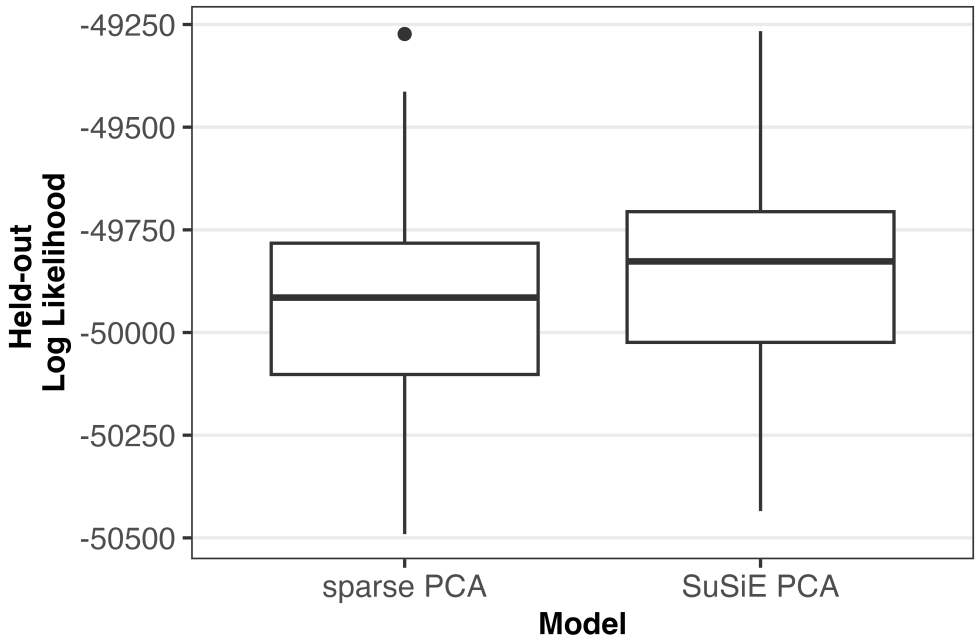

Figure S5. Held-out log-likelihood in simulations. SuSiE PCA exhibits consistently higher log-likelihood compared with those from sparse PCA on held-out data in simulations, related to Figure 2.

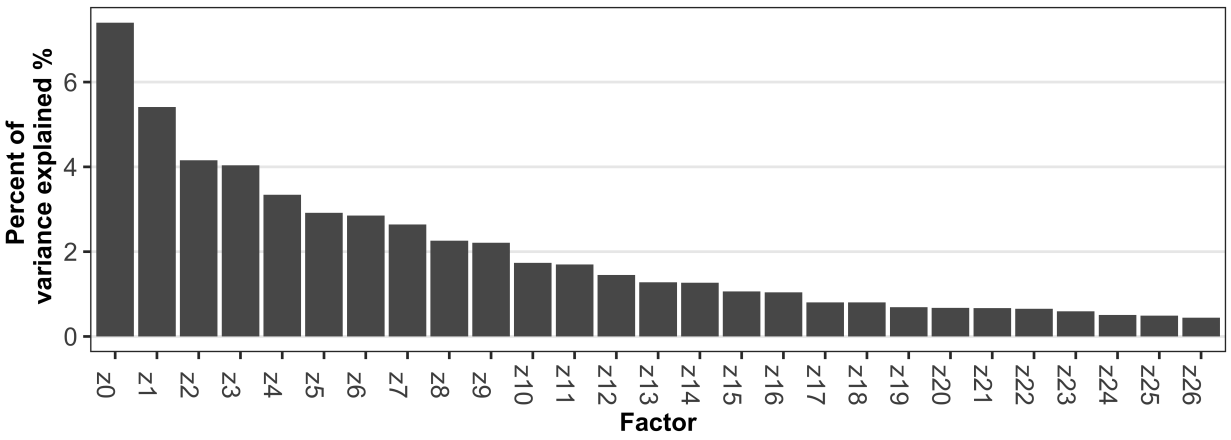

Figure S6. Percent of variance (PVE) explained by 27 factors from SuSiE PCA in GTEx z score summary data, related to Figure 3.
PVE is a measurement of variance explained by the model and is computed based on equation (3.1)

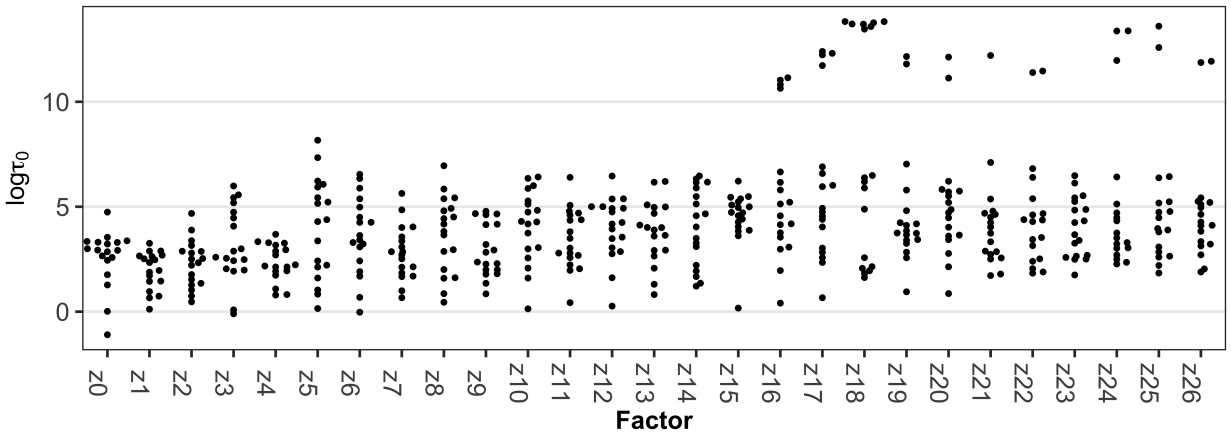

Figure S7. The MLE of the precision parameter $\log \tau_{0kl}$ in SuSiE PCA will be extremely large for those over-specified single effects in GTEx Z score summary data, related to Figure 3.
The τ_{0kl} is the inverse variance of the random variable w_{kl}. When there are excessive number of single effects specified in the SuSiE PCA, the MLE of corresponded τ_{0kl} will become extremely large and as a result shrink those redundant single effects to 0.

Figure S8. Posterior weights by factors from SuSiE PCA across different tissues in GTEx Z score summary data, related to Figure 3.
The posterior weights (or loadings) refers to the strengthen of association of the tissues contribution to the factor. The L is set to be 18 which means each factor has at most 18 tissues with non-zero effects.

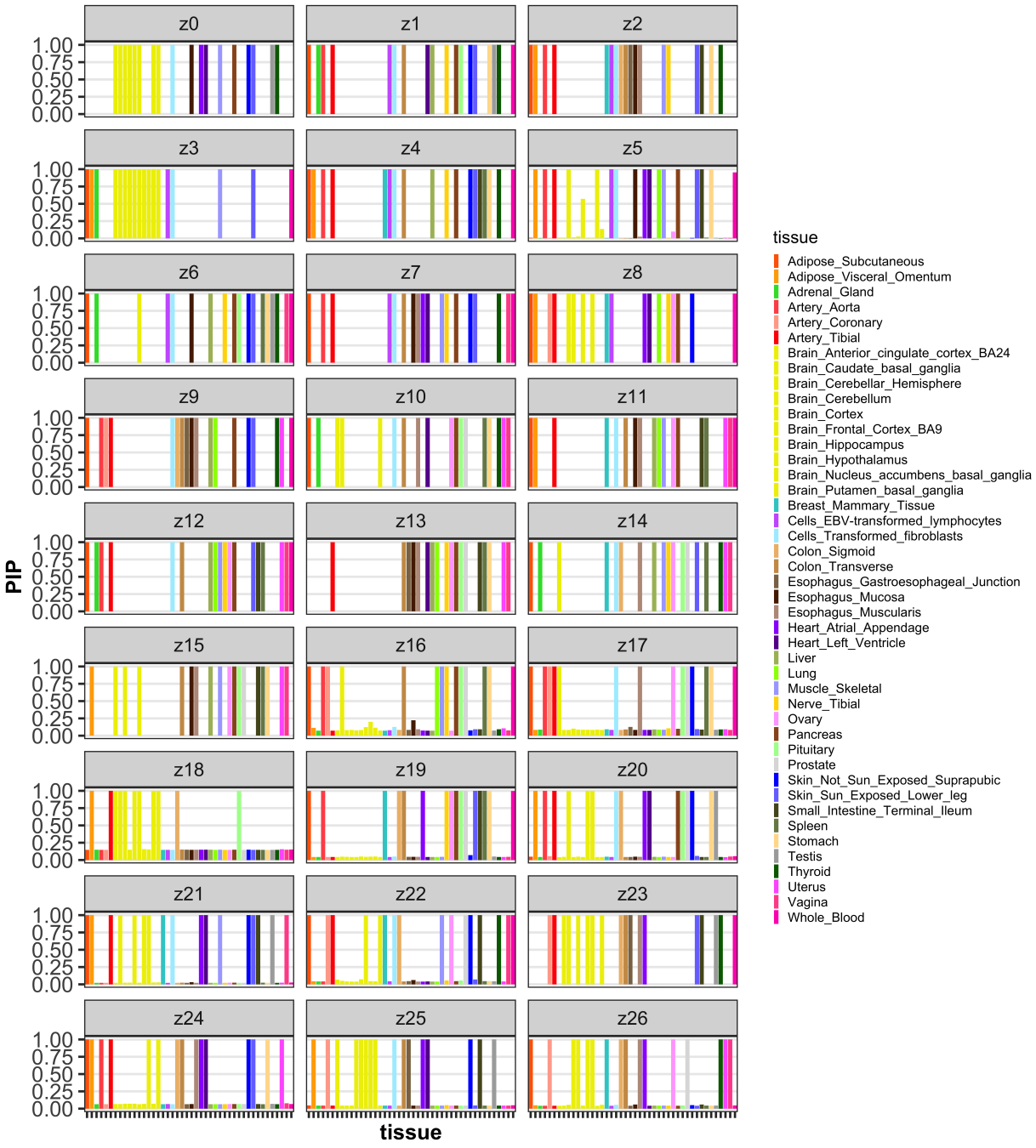

Figure S9. Posterior inclusion probabilities (PIPs) by factors from SuSiE PCA across different tissues in GTEx Z score summary data, related to Figure 3.
Most of (PIPs) are exactly 1 across different tissue by factors, implying the model is quite confident in terms of the tissues contributing to each factor

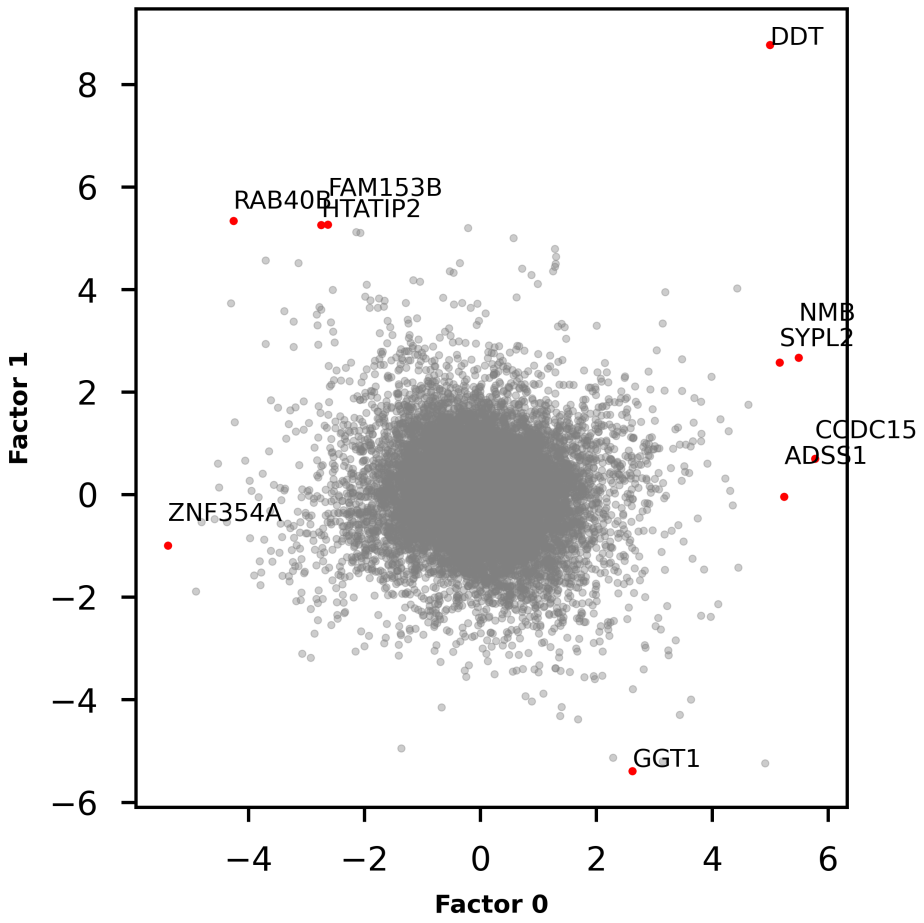

Figure S10. Scatterplot of latent factor values of z_1 vs. z_0 in GTEx Z score summary data, related to Figure 3.
Latent factor values refer to the posterior means of latent factor **Z**. Each point represents a specific gene, the genes with the top 5 absolute largest latent factor values are the red points with labels. The "outlier" gene DDT is found to be associated with testicular cancer.

Figure S11. The loading from sparse PCA in GTEx z-score data, which is similar to what we observed in SuSiE PCA, related to Figure 3.

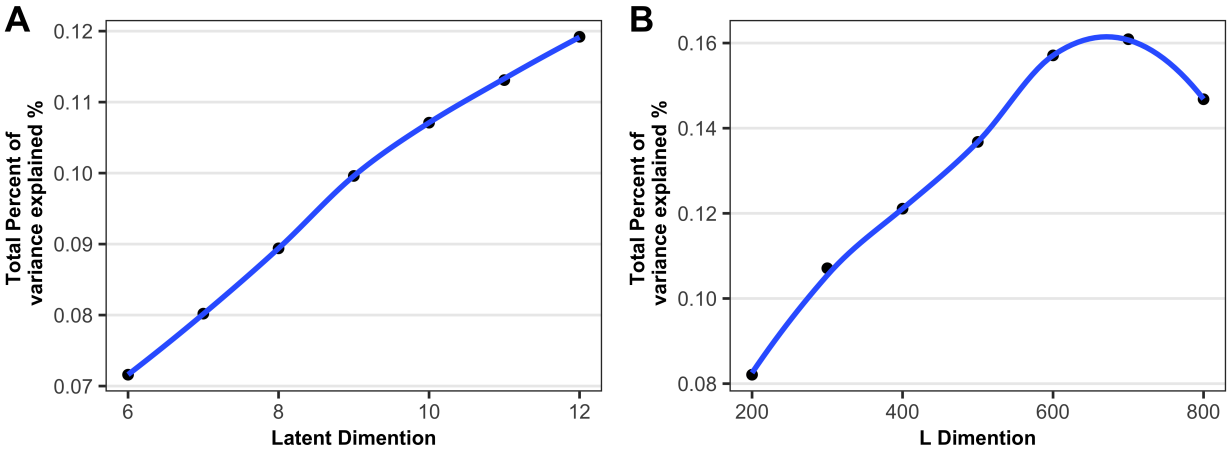

Figure S12. Total percent of variance explained (PVE) as a function of the number of latent dimension K (A) and the number of single effects L (B) in SuSiE PCA, related to Figure 4. (A) We first fixed L = 300, and varied K from 6 to 12. The increased amount between two consecutive K in total PVE becomes smaller after K reaches 10. (B) We then fixed K=10 and varied L from 200 to 800. Although the total PVE reaches its maximum at L=700, we noticed that only the first three components have 600 of downstream genes with PIP $>$ 0.9, while the rest of the components only have 200-400 genes with PIP $>$ 0.9. We then compared the results between L=300 and L=700 and realized the smaller L retains the same top significant downstream genes relevant to the component. Considering the parsimony and interpretation of the model, we finally choose K=10 and L=300.

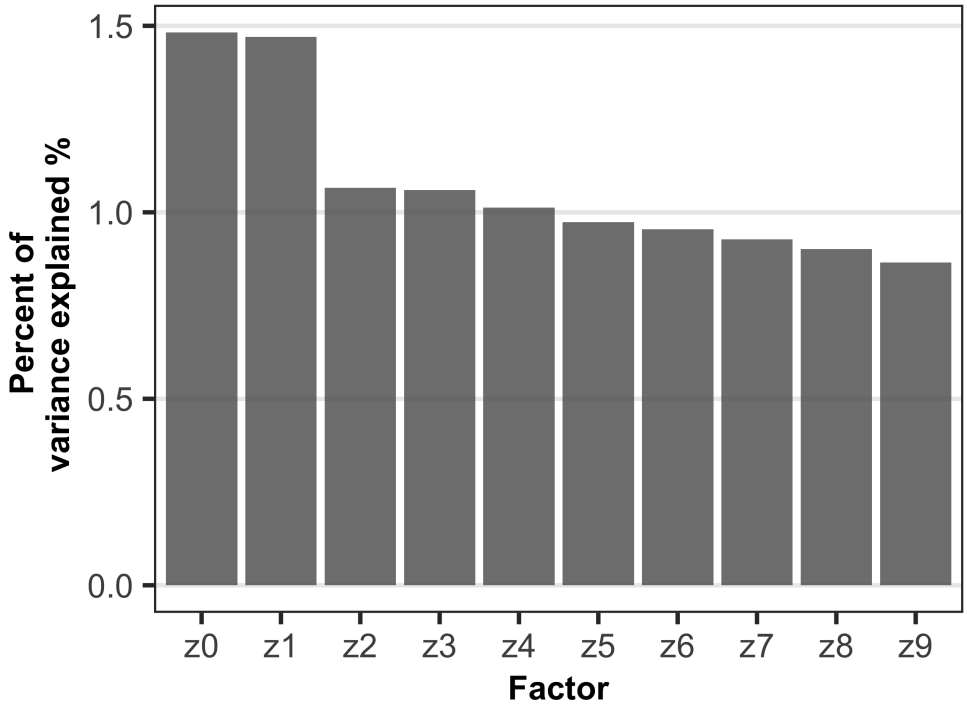

Figure S13. Percent of variance (PVE) explained by 10 factors from SuSiE PCA in gene expression data from perturb-seq data, related to Figure 4.
PVE is a measurement of variance explained by the model and is computed based on equation (3.1)

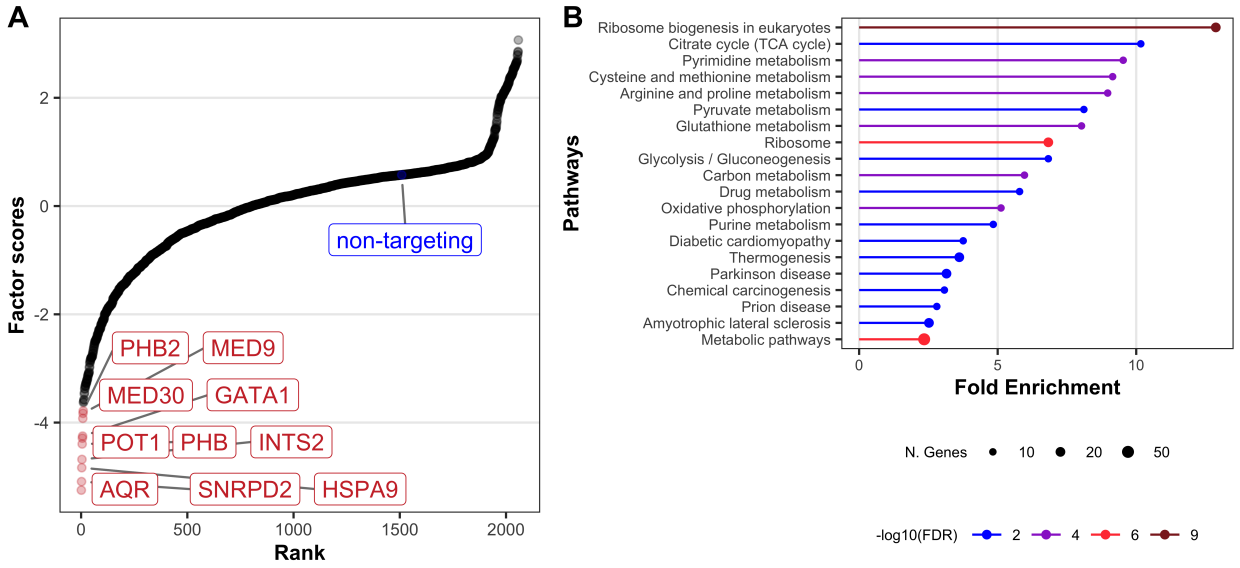

Figure S14. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 1 in perturb-seq data, related to Figure 4.

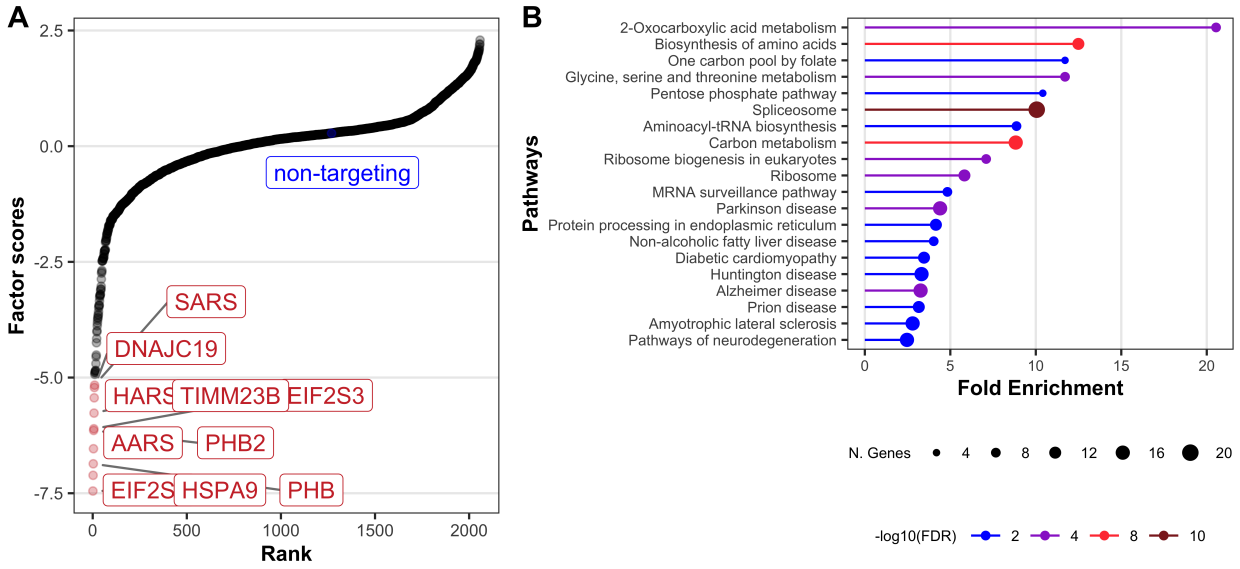

Figure S15. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 2 in perturb-seq data, related to Figure 4.

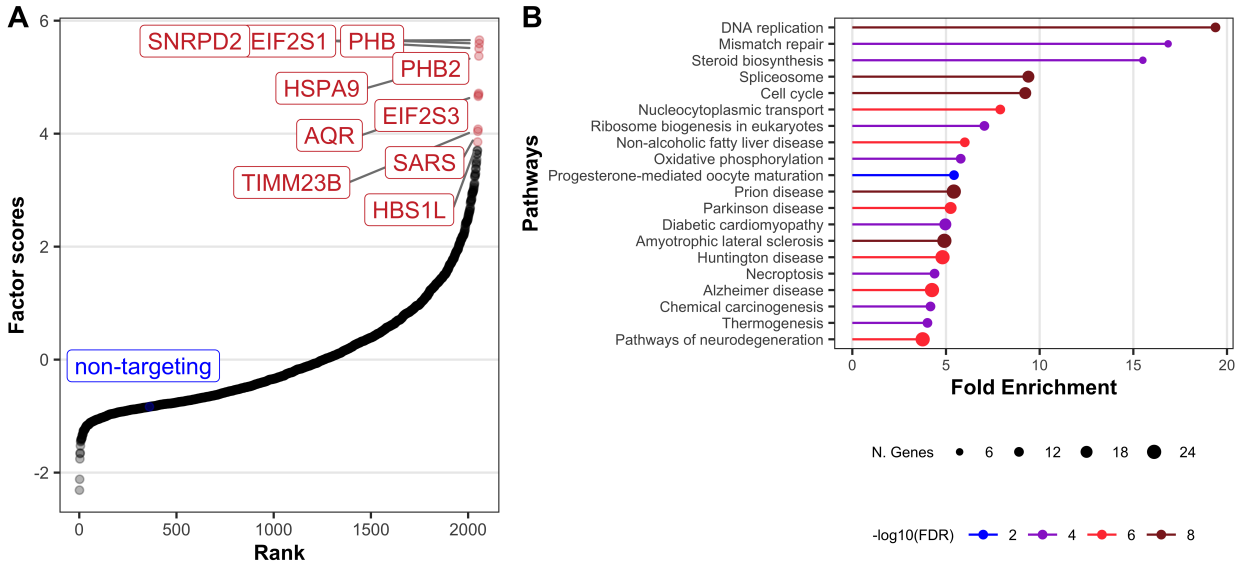

Figure S16. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 3 in perturb-seq data, related to Figure 4.

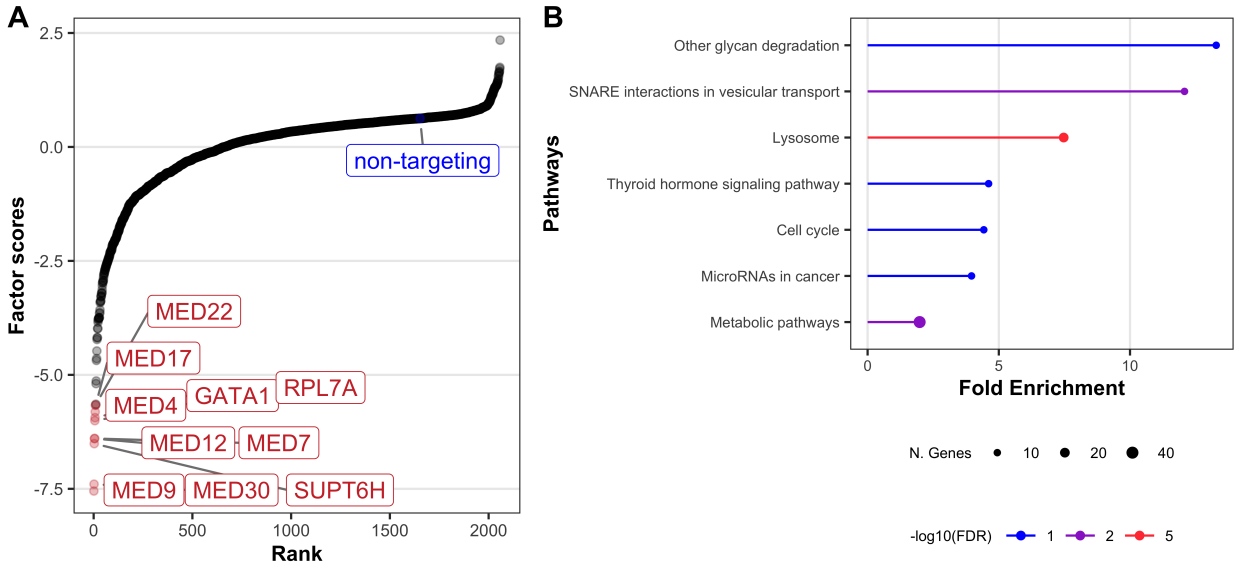

Figure S17. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 4 in perturb-seq data, related to Figure 4.

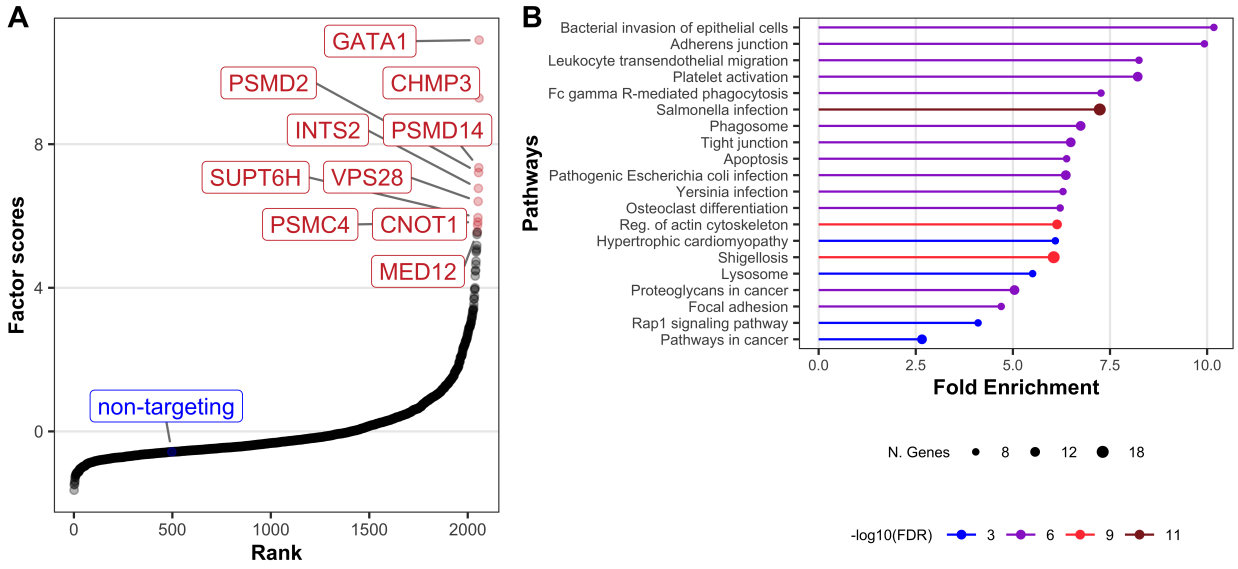

Figure S18. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 5 in perturb-seq data, related to Figure 4.

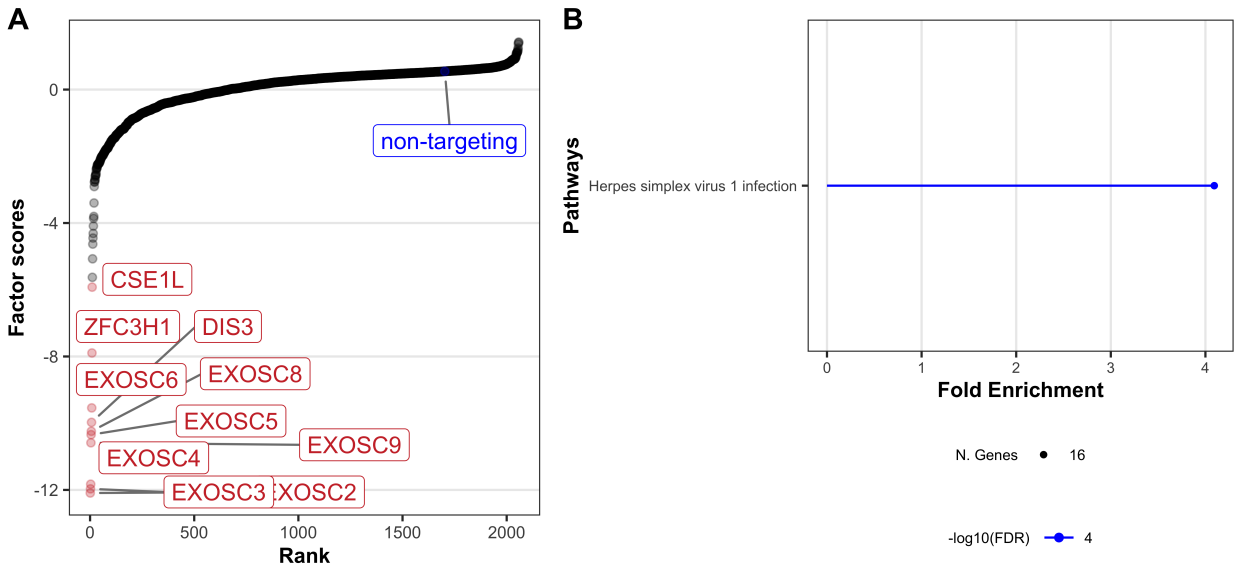

Figure S19. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 6 in perturb-seq data, related to Figure 4.

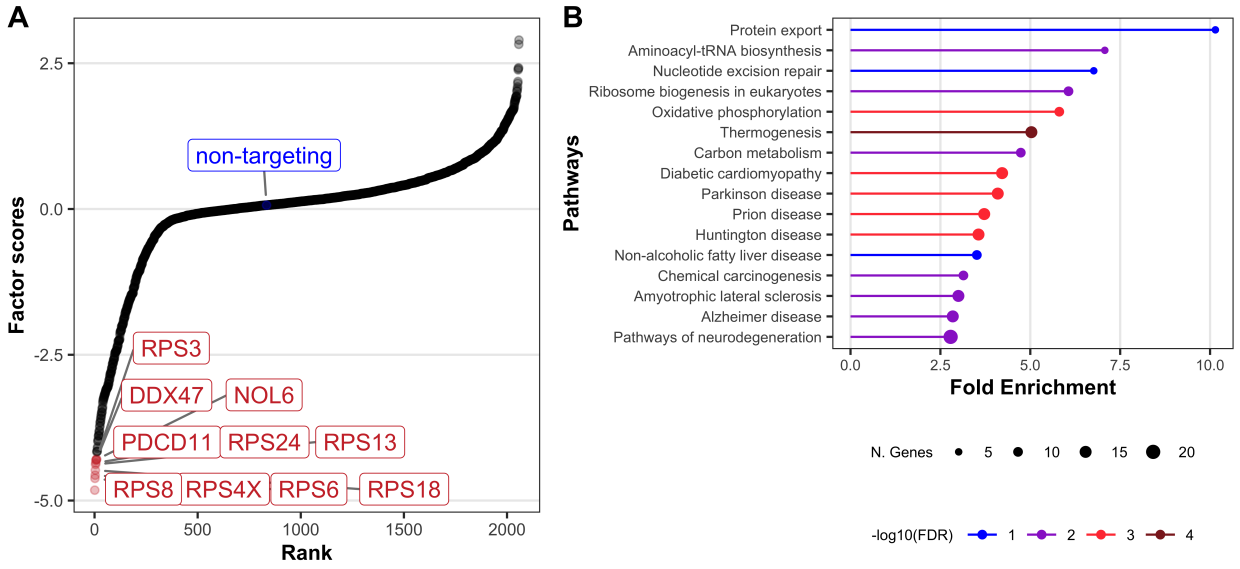

Figure S20. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 7 in perturb-seq data, related to Figure 4.

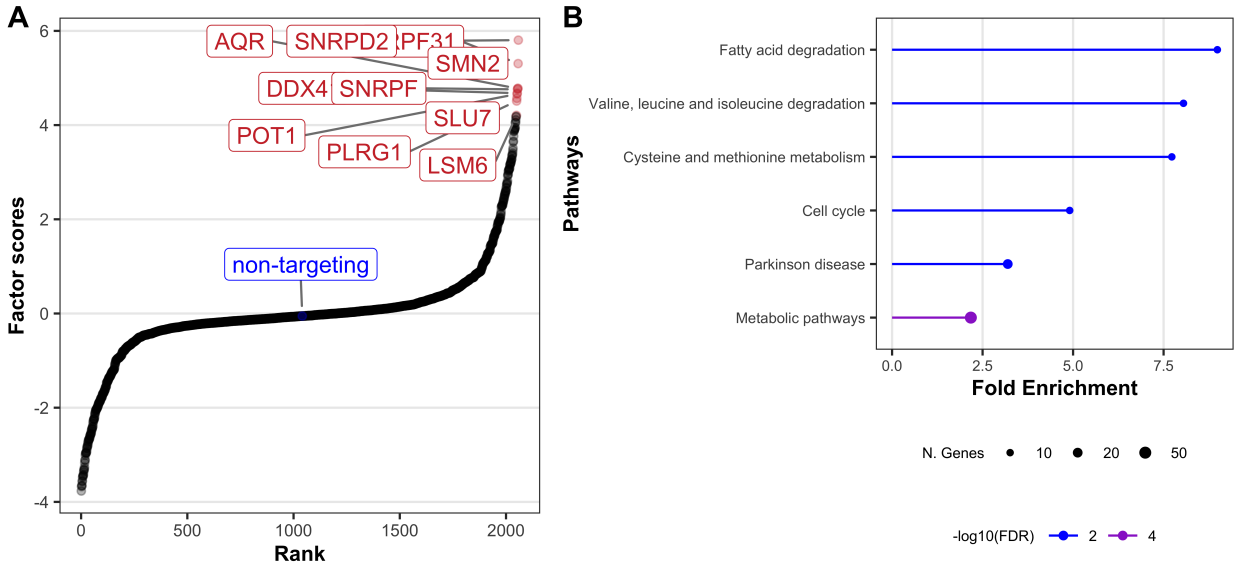

Figure S21. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 8 in perturb-seq data, related to Figure 4.

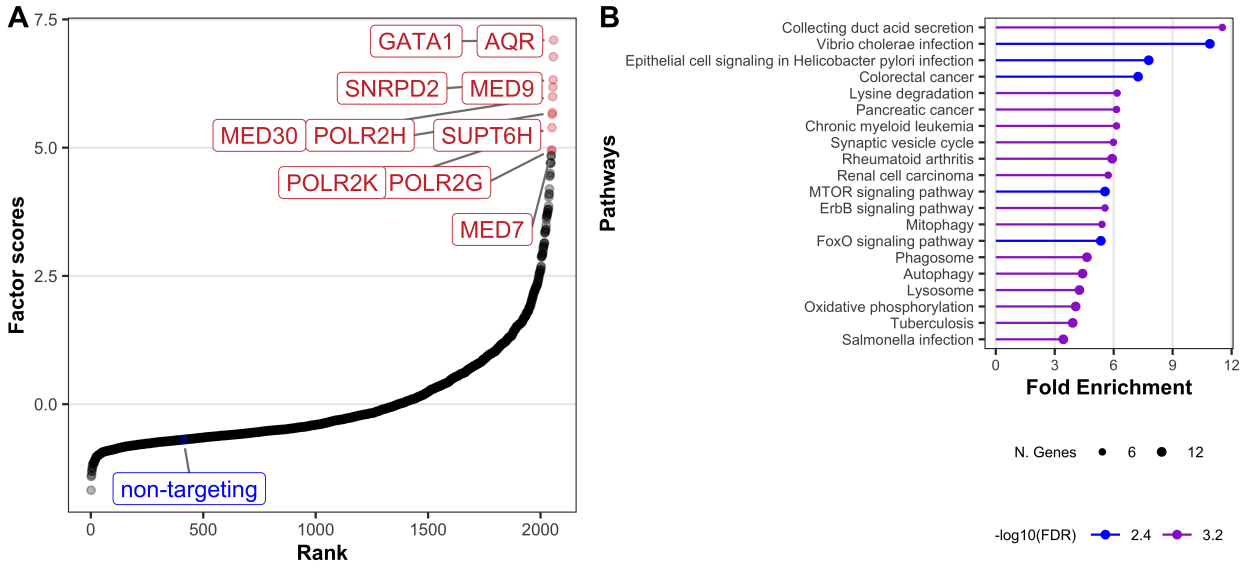

Figure S22. factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 9 in perturb-seq data, related to Figure 4.

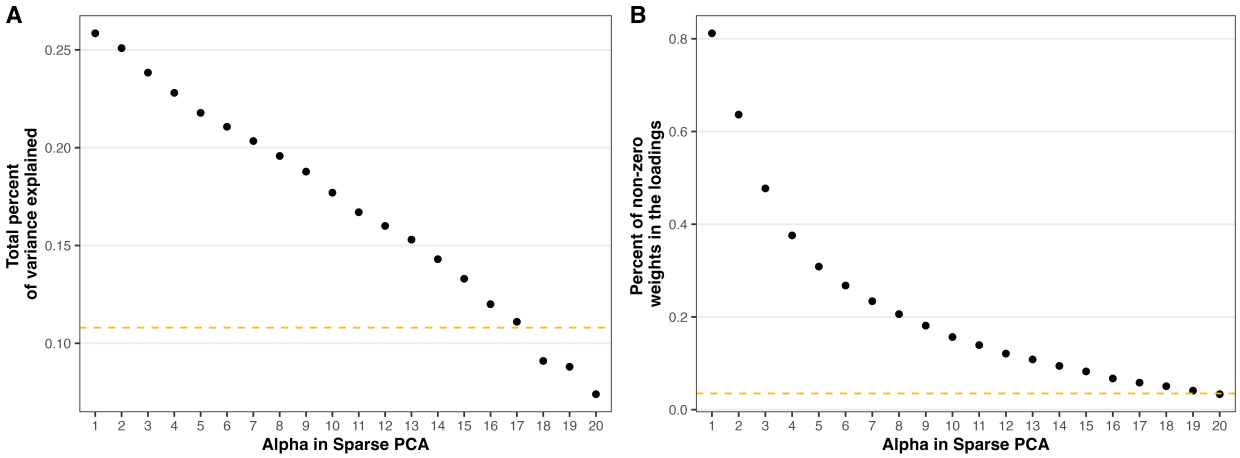

Figure S23. (A)The total percent of variance explained in sparse PCA from Perturb-seq data keeps decreasing as alpha (sparsity parameter) increases. (B) The percent of non-zero weights in the loading matrix (or sparsity) decreases as alpha increase. The yellow dashed line represents the total PVE(A) and sparsity (B) from SuSiE PCA when L = 300, related to Figure 4.

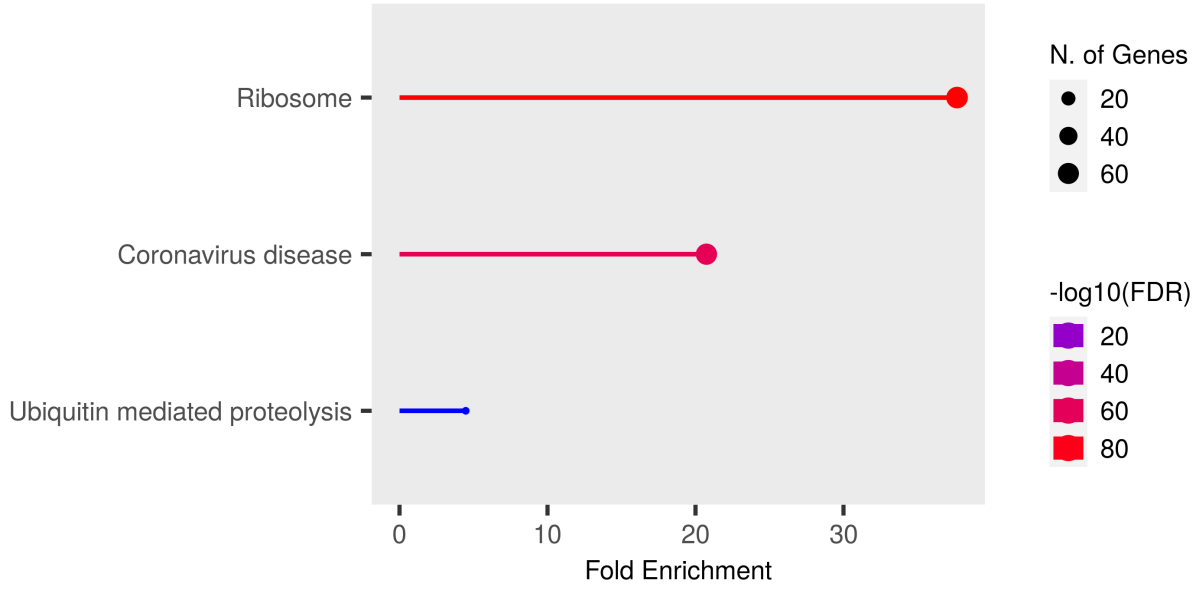

Figure S24. Gene set enrichment analysis results from the top factor in sparse PCA with alpha = 17. The enriched pathways have a similar significance level with SuSiE PCA, related to Figure 4.