Technical Report

Additional simulations for assessing FDR estimates

The simulation data used in the paper were generated by resampling the residual vectors $\boldsymbol{\epsilon}_{imk} = (\epsilon_{1imk}, \ldots, \epsilon_{Gimk})^T$. In each real data example, we obtained $\boldsymbol{\epsilon}_{imk}$ from all samples (i, m, k) and removed zero vectors (which can occur when the dataset m and condition i has only one sample, i.e., $K_{im} = 1$). This created a pool of residual vectors for each data example (i.e., MYC, Promoter, ASB). Simulation was done by randomly drawing vectors $\boldsymbol{\epsilon}_{imk}$ from this pool.

We also performed another type of simulation in which we tried to keep the dataset- and/or condition-specific characteristics as much as possible. In this simulation, to generate samples in dataset m and condition i, we first obtained all residual vectors $\boldsymbol{\epsilon}_{imk}$ from the corresponding dataset m and condition i in the original data. We then resampled these residual vectors to create K_{im} initial bootstrap samples. If a condition i in dataset m has only one sample, the residual vector will be zero. These zero residual vectors were not used in the resampling. We instead drew a non-zero residual vector from the other condition in the same dataset. If both conditions in dataset m have only one sample, the dataset will not have non-zero residual vectors. In that case, we implemented the resampling by randomly drawing non-zero residual vectors from other datasets. Since the number of replicates within each condition is small ($K_{im} = 2$ for most cases), the number of possible resamplings within each dataset m and condition i is small. In addition, when $K_{im} = 2$, the two residual vectors in dataset m and condition i only differ in a $\pm$ sign. As a result, among the four possible resampling configurations – (ϵ, ϵ), $(-\epsilon, -\epsilon)$, $(\epsilon, -\epsilon)$ and $(-\epsilon, \epsilon)$ – two configurations will have zero variance. To avoid discreteness resulted from this, for all initially sampled residual vectors, we applied an additional resampling step to shuffle the residuals within each sample. This was done as follows. Given a sampled residual vector, we went through all loci one-by-one. For locus g', we first identified 10 loci whose absolute binding in dataset m and condition i, $\bar{x}_{gim}$, were closest to $\bar{x}_{g'im}$. We then randomly picked up one locus from the 10 and used its associated residual to serve as the residual for locus g' in the final simulation data. After resampling the residuals, we then added simulated signals to condition 1, similar to what was done in the paper. The procedure described above is for the non-paired sample cases. Simulations for the paired sample case (i.e., ASB) can be performed similarly after slight modification.

Fig. R1 compares the estimated FDR and the true FDR in this simulation. The results are very similar to **Fig. S3** in the dPCA paper. dPCA was able to provide reasonable FDR estimates when $SNR > 10$. The estimates began to become biased when SNR decreased, and they were very biased when $SNR < 5$.

Our simulations show that using the current model assumptions and method (i.e., t-distribution as the null) to compute the p-value and FDR was able to provide reasonable results in our test data.

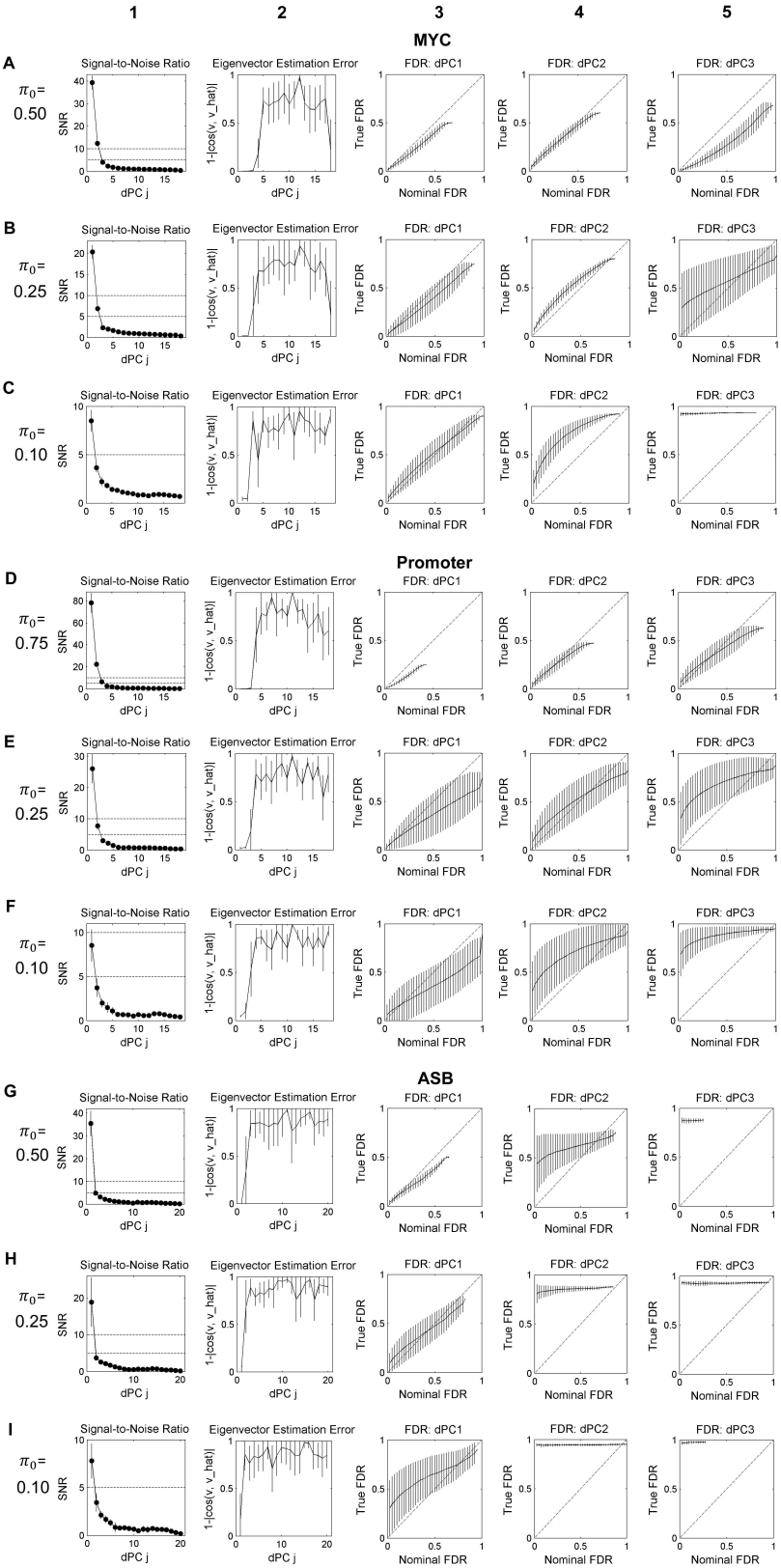

Fig. R1. Simulation results based on real data characteristics and the new resampling scheme. For each data example (MYC, Promoter, ASB), simulations were performed under different global signal-to-noise ratio (SNR) settings with the overall SNR level controlled by a parameter π_0. Increasing π_0 will increase the SNR. $\mathbf{(A)}$-$\mathbf{(C)}$: results for MYC. $\mathbf{(D)}$-$\mathbf{(F)}$: results for promoter. $\mathbf{(G)}$-$\mathbf{(I)}$: results for ASB. Each plot has 5 columns. From left to right, they are (1) estimated signal-to-noise ratio for each dPC; (2) accuracy of $\mathbf{v}_j$ estimates, measured by the cosine distance; (3)-(5) The true FDR at different levels of the estimated FDR for the first three dPCs. All plots show the average performance of 10 simulations. Vertical bars indicate ± 1 standard deviation of the 10 simulations.