total, we simulate 250 cases and 250 controls. The experiment is replicated for 100 times and the average result is provided. With the simulated data, we apply our multiple testing procedure (LIS) and three baseline procedures: the BH procedure, the adaptive p-value procedure (AP), and the local false discovery rate procedure (localFDR). Because the dependence structure is real and the ground truth parameters are unknown to us, we do not have the oracle procedure in the simulations on genetic data.

With the simulated genetic data, we use two commonly used tests in genetic association studies, namely two-proportion z-test and Cochran-Armitage’s trend test (CATT) (Cochran, 1954; Armitage, 1955; Slager & Schaid, 2001; Freidlin et al., 2002) as the individual tests for the association of each SNP. CATT also yields an asymptotic \( N(0, 1) \) under the null and \( N(\mu, 1) \) under the alternative (\( \mu \) is nonzero). Therefore, we parameterize \( \psi = (\mu_1, \sigma_1^2) \) where \( \mu_1 \) and \( \sigma_1^2 \) are the mean and variance of the test statistics under alternative.

The graph structure is built as follows. Each SNP becomes a node in the graph. For each SNP, we connect it with the SNP with the highest \( r^2 \) value with it. There are in total 490 edges in the graph. We further categorize the edges into a high correlation edge set \( E_h \) (\( r^2 \) above 0.8), medium correlation edge set \( E_m \) (\( r^2 \) between 0.5 and 0.8) and low correlation edge set \( E_l \) (\( r^2 \) between 0.25 and 0.5). We have three different parameters (\( \phi_h, \phi_m, \) and \( \phi_l \)) for the three sets of edges. Then the density of \( \theta \) in formula (1) takes the form

\[
P(\theta, \phi) \propto \exp\left\{ \sum_{(i,j) \in E_h} \phi_h I(\theta_i = \theta_j) + \sum_{(i,j) \in E_m} \phi_m I(\theta_i = \theta_j) + \sum_{(i,j) \in E_l} \phi_l I(\theta_i = \theta_j) \right\},
\]

where \( I(\theta_i = \theta_j) \) is an indicator variable that indicates whether \( \theta_i \) and \( \theta_j \) take the same value. In the MCMC algorithm, we run the Markov chain for 20,000 iterations with a burn-in of 100 iterations. In the PCD algorithm, we generate 100 particles. In each iteration of PCD learning, the particles move forward for 5 iterations (the \( n \) parameter in PCD-\( n \)). The learning rate in PCD gradually decreases as suggested by Tieleman (2008). The EM algorithm converges after about 10 to 20 iterations, which usually take less than 10 minutes on a 3.00GHz CPU.

Figure 5 shows the performance of the procedures in the additive models with the homozygous relative risk set to 1.2 and 1.3. The test statistics are from a two-proportion z-test. We have also replicated the simulations on Cochran-Armitage’s trend test, and the results are almost the same. In Figure 5, table (1) summarizes the empirical FDR and the total number of true positives (\( \#TP \)) of our LIS procedure, BH, AP and localFDR (localFDR) in the additive models with different (homozygous) relative risk levels, when we vary \( \tau \) and when we vary the nominal FDR level \( \alpha \). We regard a SNP having \( r^2 \) above \( \tau \) with any causal SNP as an associated SNP, and we regard a rejection of the null hypothesis for an associated SNP as a true positive. Our LIS procedure and localFDR are valid while being conservative. BH and AP appear liberal in some of the configurations. In any of the circumstances, our LIS procedure can identify more associated SNPs than the baselines. We can find a clue to why our procedure LIS is being conservative from the results in Figure 3. In basic simulation 2, we observe that when the parameters \( \mu \) and \( \phi \) are heterogeneous and we carry out the data-driven procedure under the homogeneous parameter assumption, the data-driven procedure is conservative. The discrepancy between the nominal FDR level and the empirical FDR level increases as the parameters move further away from homogeneity. Although we assign three different parameters \( \phi_h, \phi_m, \) and \( \phi_l \) to \( E_h, E_m \) and \( E_l \) respectively, the edges within the same set (e.g. \( E_l \)) may still be heterogeneous. The fact that the LIS procedure captures more true positives than the baselines while remaining more conservative in many configurations indicates that the local indices of significance provide a ranking more efficient than the ranking provided by the \( p \)-values from the individual tests. Therefore, we further plot the ROC curves and precision-recall (PR) curves when we rank SNPs by LIS and by the \( p \)-values from the two-proportion z-test. The ROC curve and PR curve are vertically averaged from 100 replications. Subfigures (2a)-(2f) are for the additive model with homozygous relative risk level set to 1.2. Subfigures (3a)-(3f) are for the additive model with homozygous relative risk level set to 1.3. It is observed that the curves from LIS dominate those from the \( p \)-values from individual tests in most places, which further suggests that LIS provides a more efficient ranking of the SNPs than the individual tests.

Figure 6 shows the performance of the procedures in the dominant model and the recessive model with the homozygous relative risk set to be 1.2. The test statistics are from a two-proportion z-test. In Figure 6, table (1) summarizes the empirical FDR and the total number of true positives (\( \#TP \)) of our LIS procedure, BH, AP and localFDR (localFDR) in the dominant model and the recessive model when we vary \( \tau \) and when we vary the nominal FDR level \( \alpha \). Our LIS procedure and localFDR are valid while being conservative in all configurations, and they appear more conservative in the recessive model than in the dominant model. On the other hand, BH and AP appear liberal in the recessive model. Our LIS procedure still confers an advantage