

Supplementary information

A computational model for sample dependence in hypothesis testing of genomic data

No	Phenotype	No	Phenotype
1	Flowering time at Arkansas	16	Seed length
2	Flowering time at Faridpur	17	Seed width
3	Flowering time at Aberdeen	18	Seed volume
4	FT ratio of Arkansas Aberdeen	19	Seed surface area
5	FT ratio of Faridpur Aberdeen	20	Brown rice seed length
6	Culm habit	21	Brown rice seed width
7	Flag leaf length	22	Brown rice surface area
8	Flag leaf width	23	Brown rice volume
9	Panicle number per plant	24	Seed length width ratio
10	Plant height	25	Brown rice length width ratio
11	Panicle length	26	Straight head susceptibility
12	Primary panicle branch number	27	Blast resistance
13	Seed number per panicle	28	Amylose content
14	Florets per panicle	29	Alkali spreading value
15	Panicle fertility	30	Protein content

Table S1. A list of 30 quantitative traits in the 44K_SNP data used in this study.

No	Phenotype
1	Leaf: intensity of green color
2	Leaf blade: pubescence of blade surface
3	Leaf: the shape of ligule
4	Flag leaf: attitude of the blade(early observation)
5	Flag leaf: attitude of the blade(late observation)
6	Culm: habit
7	Panicle: awns
8	Spikelet: color of the tip of lemma
9	Panicle: attitude in relation to stem
10	Panicle: attitude of secondary branching
11	Panicle: attitude of branch
12	Panicle: exsertion
13	Lemma: color
14	Glume: length

Table S2. A list of 14 qualitative traits in the 580K_KNU data used in this study.

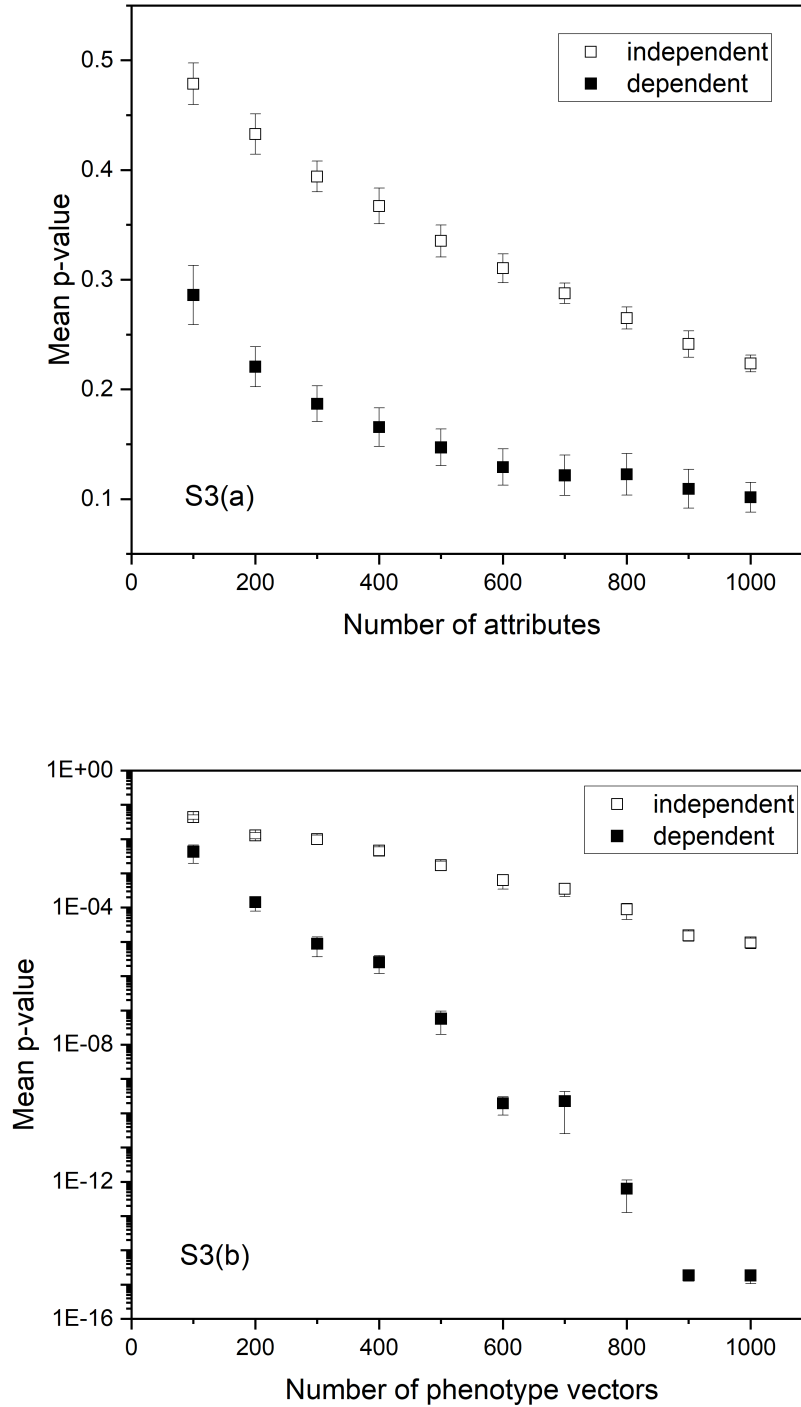


Figure S3. (a) Plots of the mean p value obtained by the Kendall correlation test versus the number of attributes using the independent ($\square$) and dependent ($\blacksquare$) data. The mean is taken over all possible pairs of 20 variables and the error bars represent the standard error estimated over 20 independent trials. (b) Plots of the mean p value obtained by the F test versus the number of phenotype vectors using the independent ($\square$) and the dependent ($\blacksquare$) data. The mean is taken over 20 phenotype vectors and the error bars represent the standard error estimated over 20 independent trials. The other parameters used in the association data are the minor allele frequency $q = 0.3$ and the shuffling rate $\kappa = 0.4$. Note that the ordinate is on the logarithmic scale.

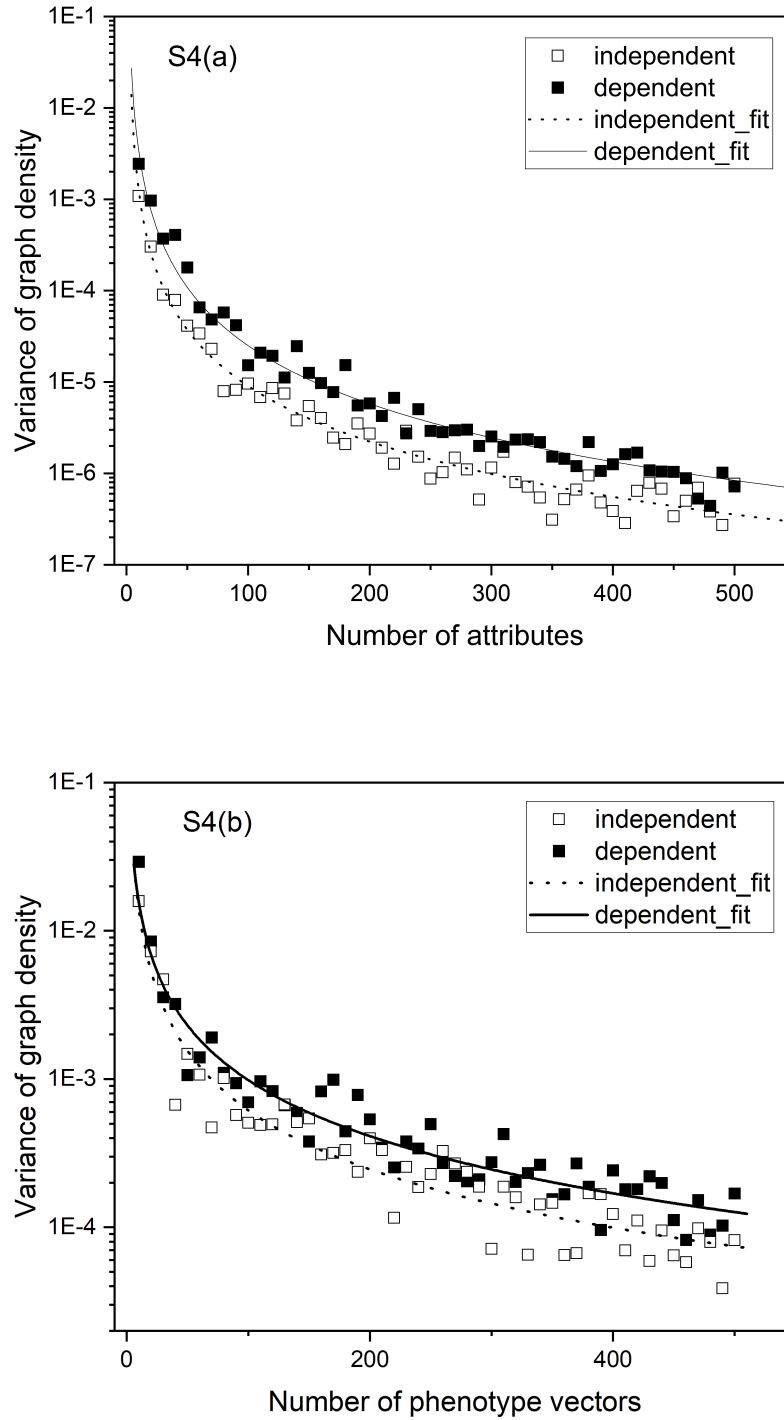


Figure S4. Plots of the variance of the graph density versus the sample size of dependent (■) and independent (□) data using the Kendall test. The solid and dashed lines are fitted to dependent and independent data, respectively, with parameters estimated. (a) Plots using model-free data fitted with $\alpha = 2.11$, $\lambda = 1.12$ for dependent data and $\alpha = 1.99$, $\lambda = 3.77$ for independent data. Each variable contains 30 attributes and 20 independent trials were run to estimate the sample variance. (b) Plots using model-based data fitted with $\alpha = 1.09$, $\lambda = 0.36$ for dependent data and $\alpha = 0.89$, $\lambda = 0.21$ for independent data. Each genotype generates 30 phenotypes and 20 independent trials were run to estimate the sample variance.

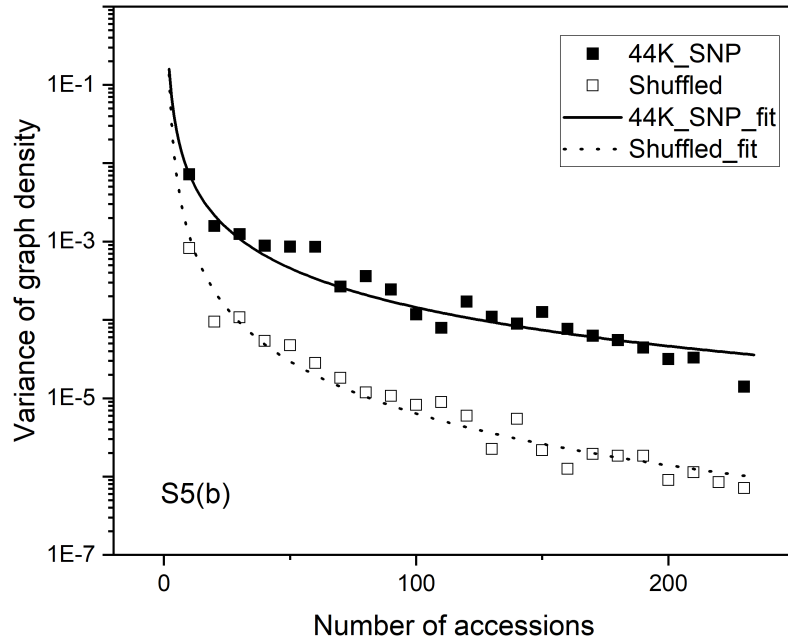
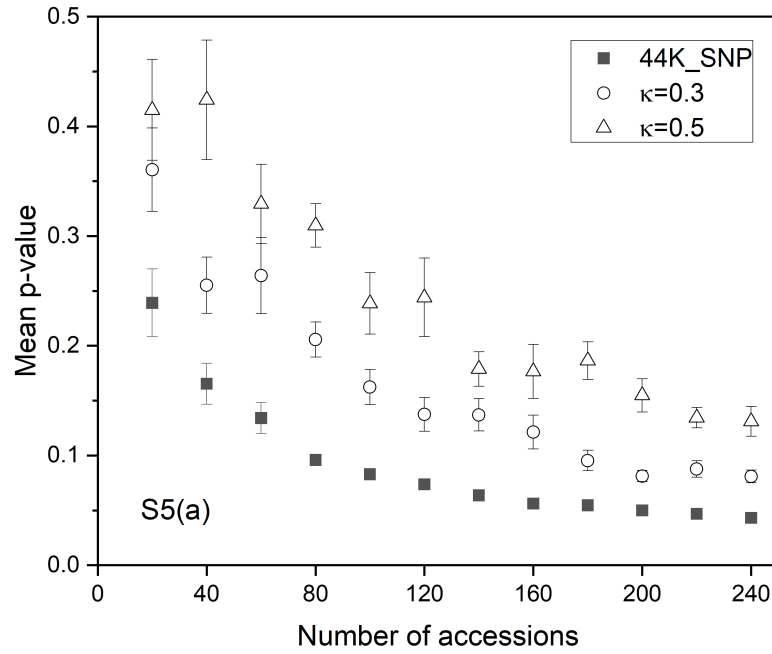


Figure S5. (a) Plots of the mean p value obtained from the F test of the 44K_SNP data using the seed volume phenotype (■) versus sample size, together with those of different shuffling rates, $\kappa = 0.3$ (○) and $\kappa = 0.5$ (△). (b) The plot of the variance of the graph density versus the sample size of the 44K_SNP data (■) and the shuffled data (□). Graph density was obtained using the Kendall correlation test. The solid and dashed lines are fitted to the 44K_SNP data and shuffled data, respectively. The estimated parameters are $\alpha = 1.61$, $\lambda = 0.74$ for the 44K_SNP data and $\alpha = 2.19$, $\lambda = 3.87$ for the shuffled data.