The American Journal of Human Genetics, Volume *80*

Supplemental Data

Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test

Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin

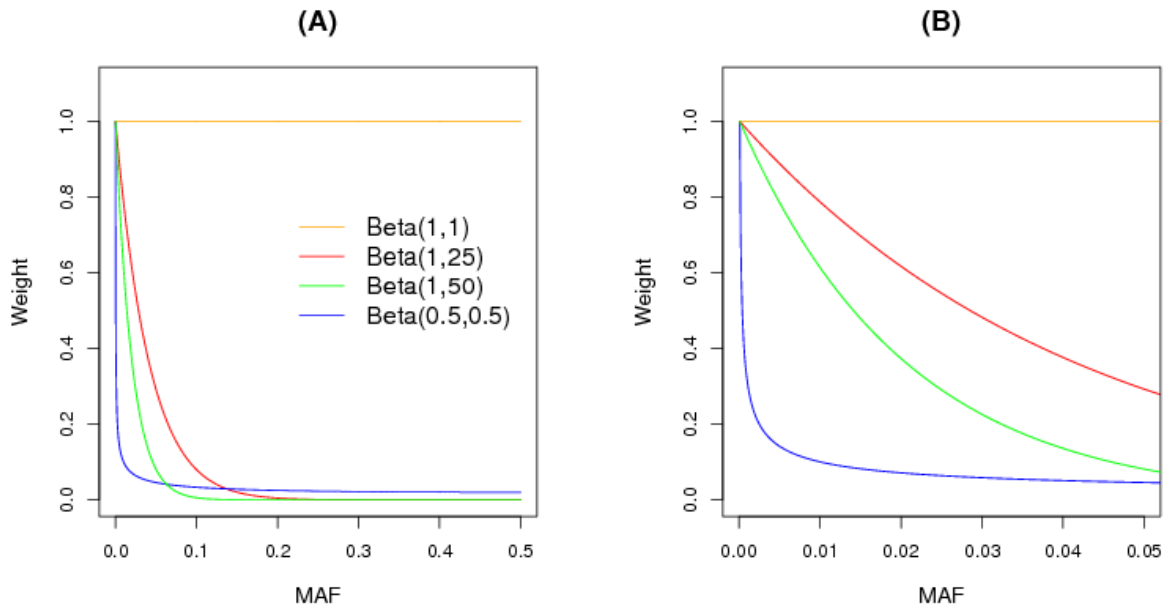

Figure S1. Examples of Flexible Weight Functions Used in SKAT

Examples of the weight function used in SKAT, where the weight is assumed to be a function of MAF using the beta density function *Beta*(MAF, a_1,a_2): *Beta*(1, 1), *Beta*(1, 25), *Beta*(1, 50), and *Beta*(0.5, 0.5). (A) and (B) represent the same functions, while (B) zooms into the region with MAF <0.05.

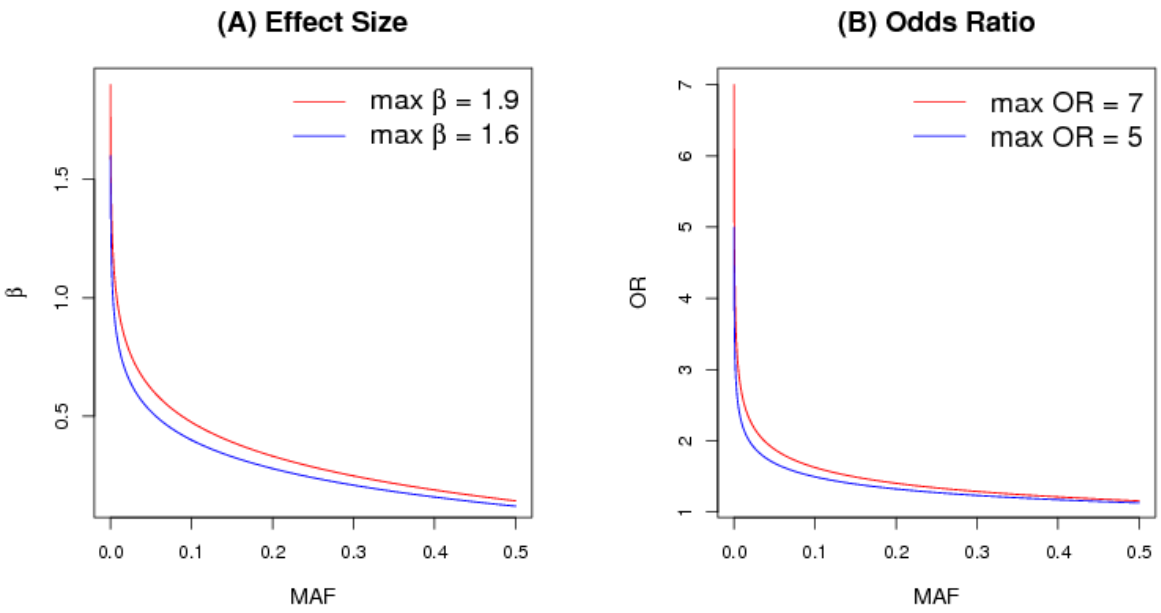

Figure S2. Effect Size/Odds Ratio Functions Used in the Simulation Studies
Effect sizes for the causal variants in the continuous trait simulations (A) and odds ratios for the causal variants in case-control dichotomous trait simulations (B) as a function of population minor allele frequency. For continuous traits, the effect size is $|\beta| = c\left|\log_{10} MAF_j\right|$ with c = 0.4 (blue) and c=0.475 (red), which correspond to maximum effect size $|\beta_j|$=1.6 and 1.9, respectively, when MAF=10^{-4}. For dichotomous traits, log odds ratio is $|\beta| = c\left|\log_{10} MAF_j\right|$ with c=ln5/4 (blue) and c=ln7/4 (red), corresponding to maximum OR = 5 ($|\beta_j|$= ln5) and 7 ($|\beta_j|$= ln7), respectively, when MAF=10^{-4}.

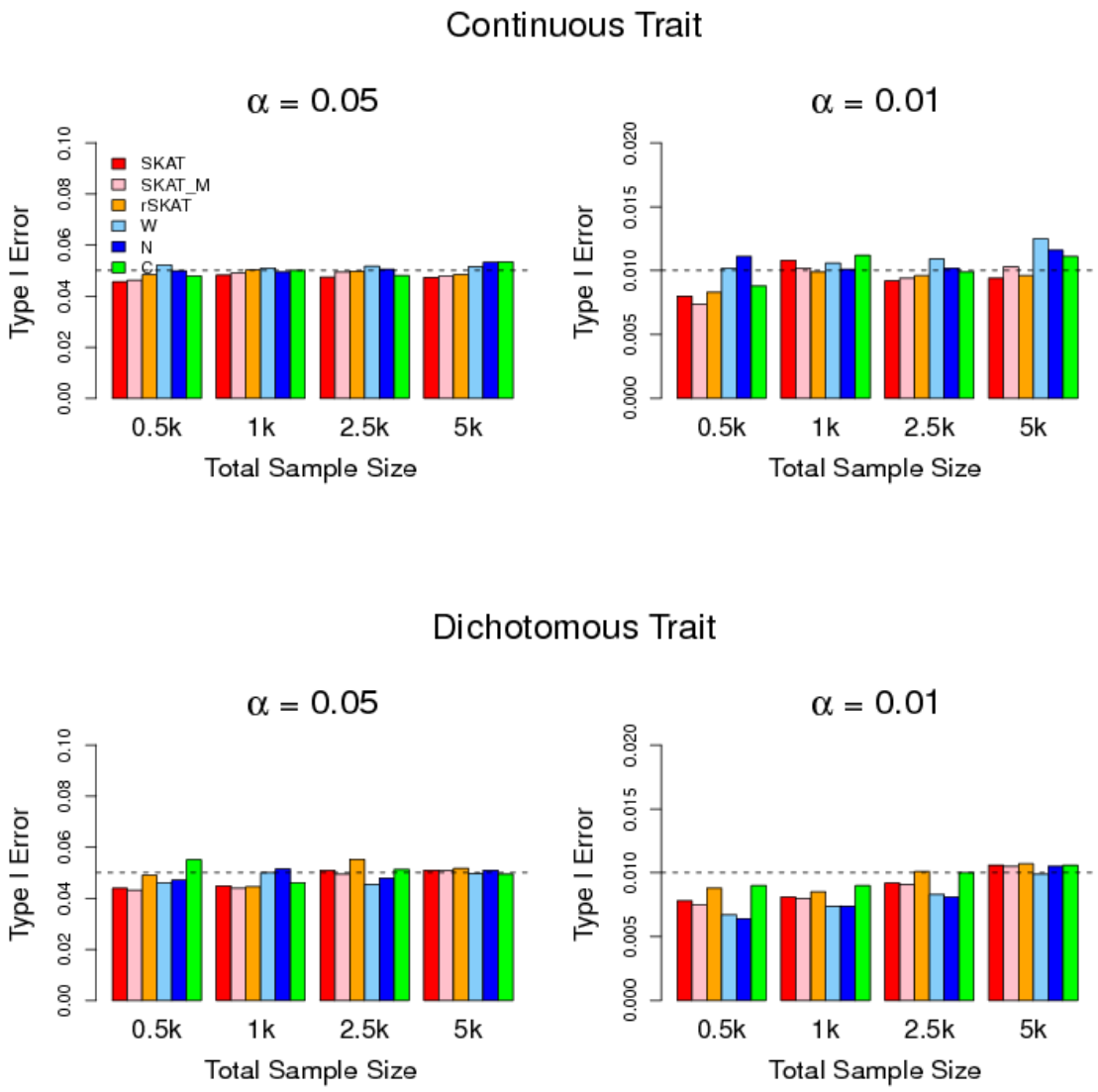

Figure S3. Type I Error Rates of SKAT and Burden Tests Using Simulation Studies
Empirical type I error rate to test an association between randomly selected 30kb regions with a continuous trait (top panel) or a case-control dichotomous trait (bottom panel) at the levels α=0.05 (left panel) and α=0.01 (right panel). Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, type I error rate was estimated as the proportion of p-values < α under the null hypothesis using 10,000 simulated data sets. For $\alpha=10^{-4}$, $10^{-5,}$ $10^{-6,}$, the empirical type I error rates of SKAT are close to the nominal values and are given in Supplementary Table 1.

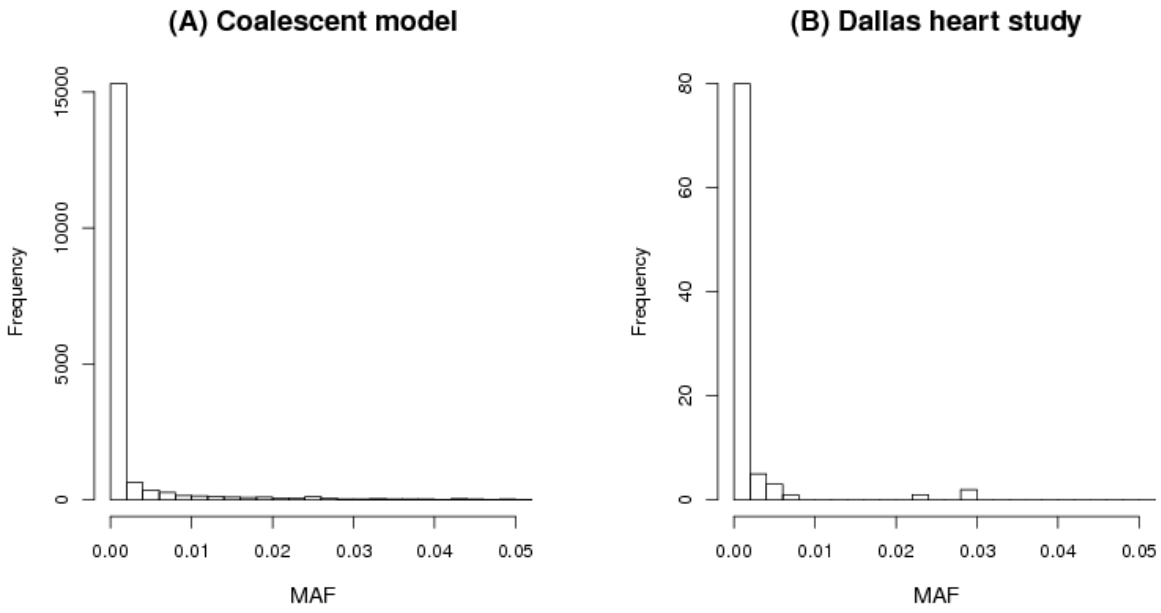

Figure S4. Distribution of Allele Frequencies for Simulated and Real Data
MAFs for simulated data are based on the population allele frequencies under the coalescent model (A) and MAFs for real data are estimated from the observed Dallas Heart study data (B).

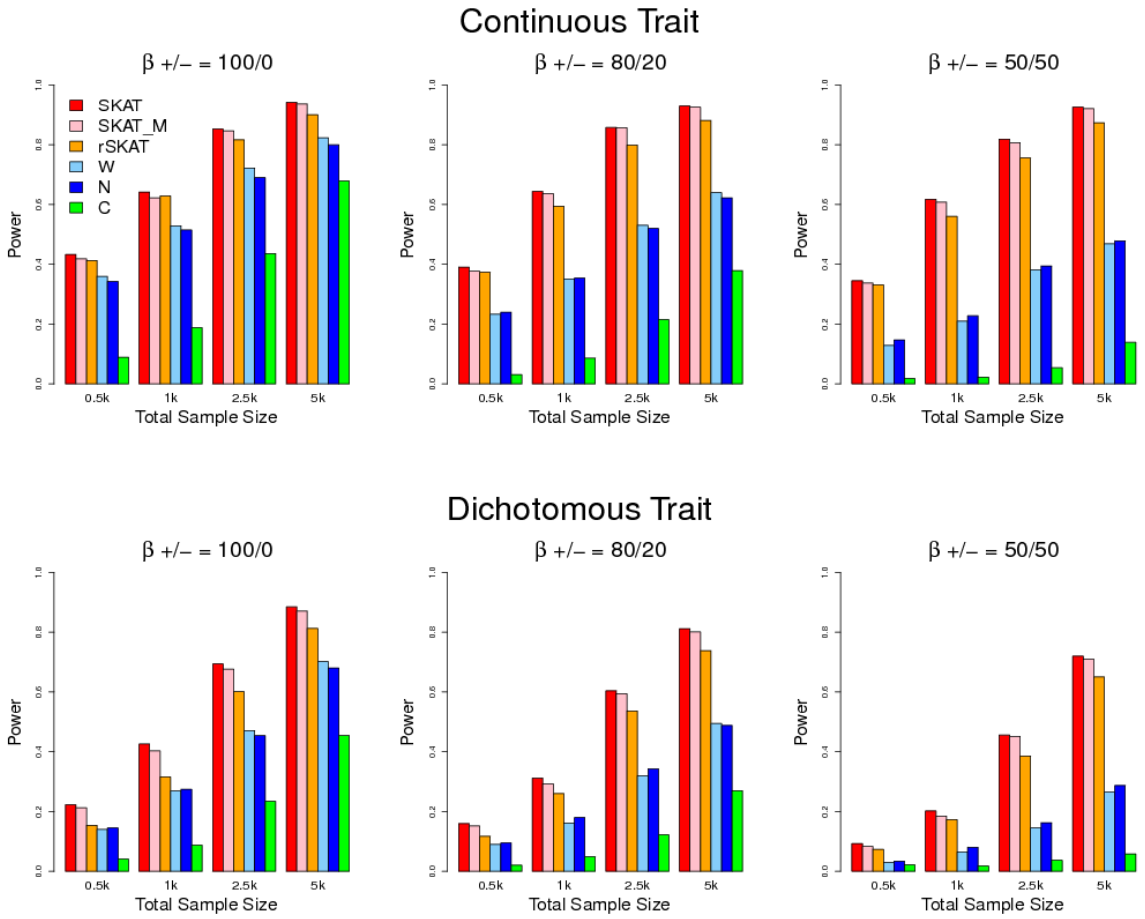

Figure S5. Power Comparisons for SKAT and Burden Tests Using Simulation Studies at the Level $\alpha=10^{-2}$

Empirical power under the same configuration as that in Figure 1 except $\alpha= 10^{-2}$. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-2}$ using 1000 simulated data sets.

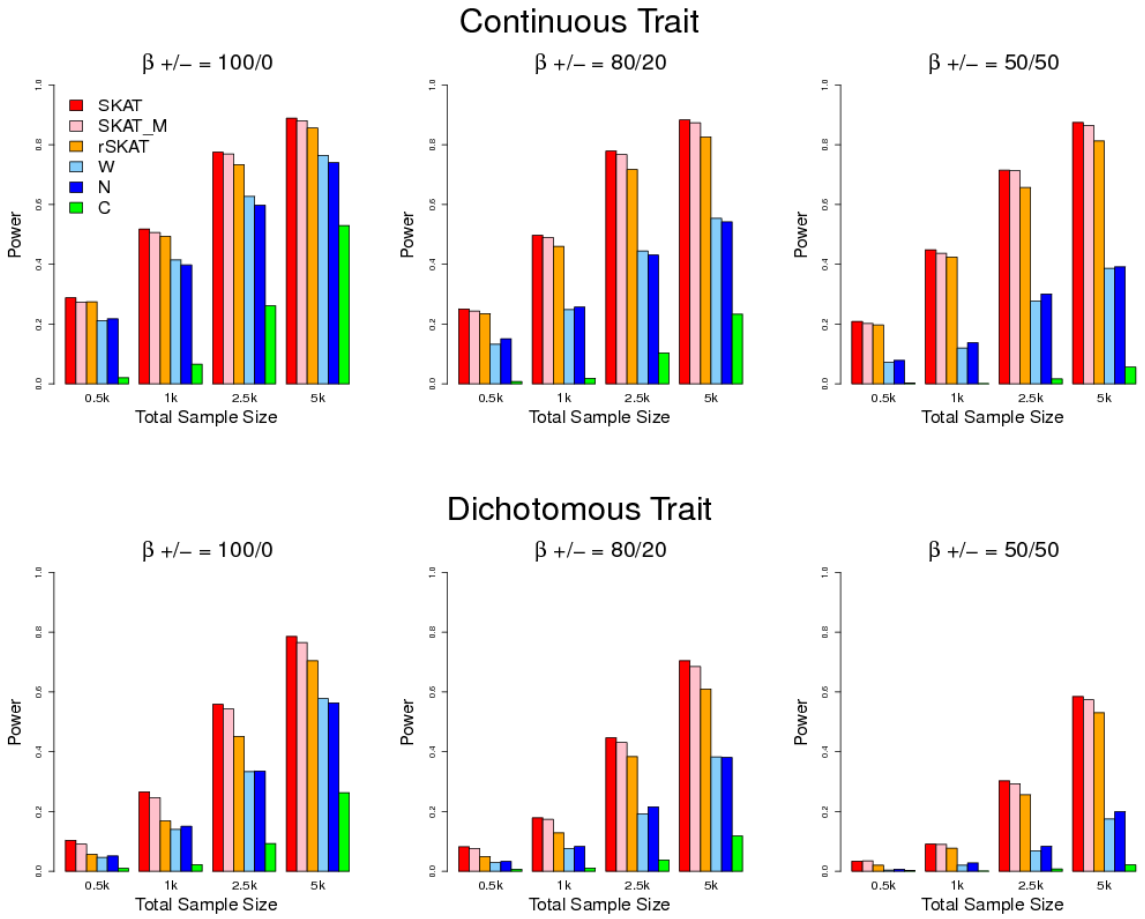

Figure S6. Power Comparisons for SKAT and Burden Tests Using Simulation Studies at the Level $\alpha = 10^{-3}$

Empirical power under the same configurations as that in Figure 1 except $\alpha=10^{-3}$. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-3}$ using 1000 simulated data sets.

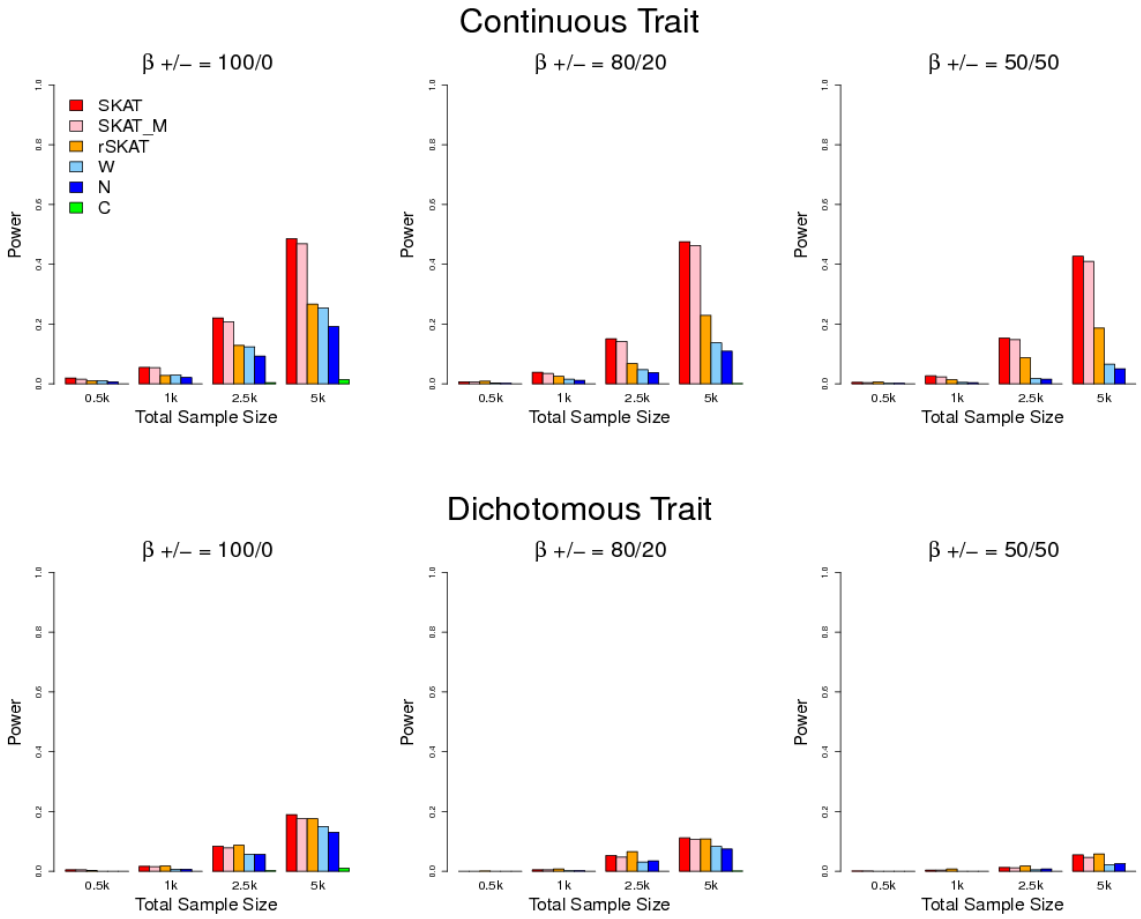

Figure S7. Power Comparisons for SKAT and Burden Tests Using Simulation Studies When Causal Variants Have MAF < 1%

Empirical power at the genome-wide $\alpha=10^{-6}$ level under the same configuration to Figure 1 except 5% of the rare variants with MAF<1% were selected as causal variants. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

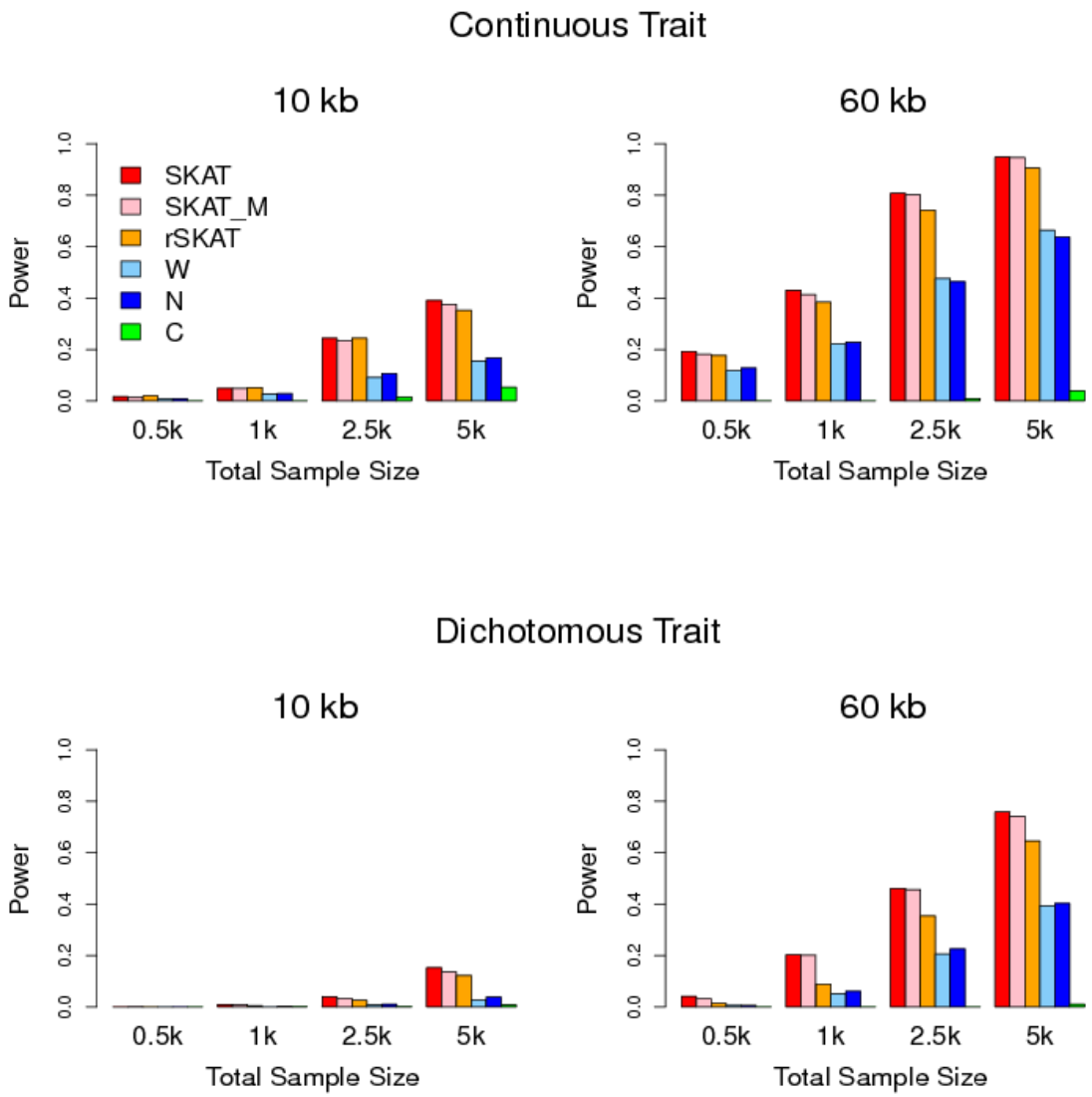

Figure S8. Power Comparisons Comparing SKAT with Burden Test Using Simulation Studies with Different Region Sizes

Empirical power at the genome-wide $\alpha=10^{-6}$ level in the same setting as that in Figure 1 except the region size is 10kb (left panel) and 60kb (right panel). From left to right, the plots consider settings in which the coefficients for the causal rare variants are 100%, positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < α using 1000 simulated data sets.

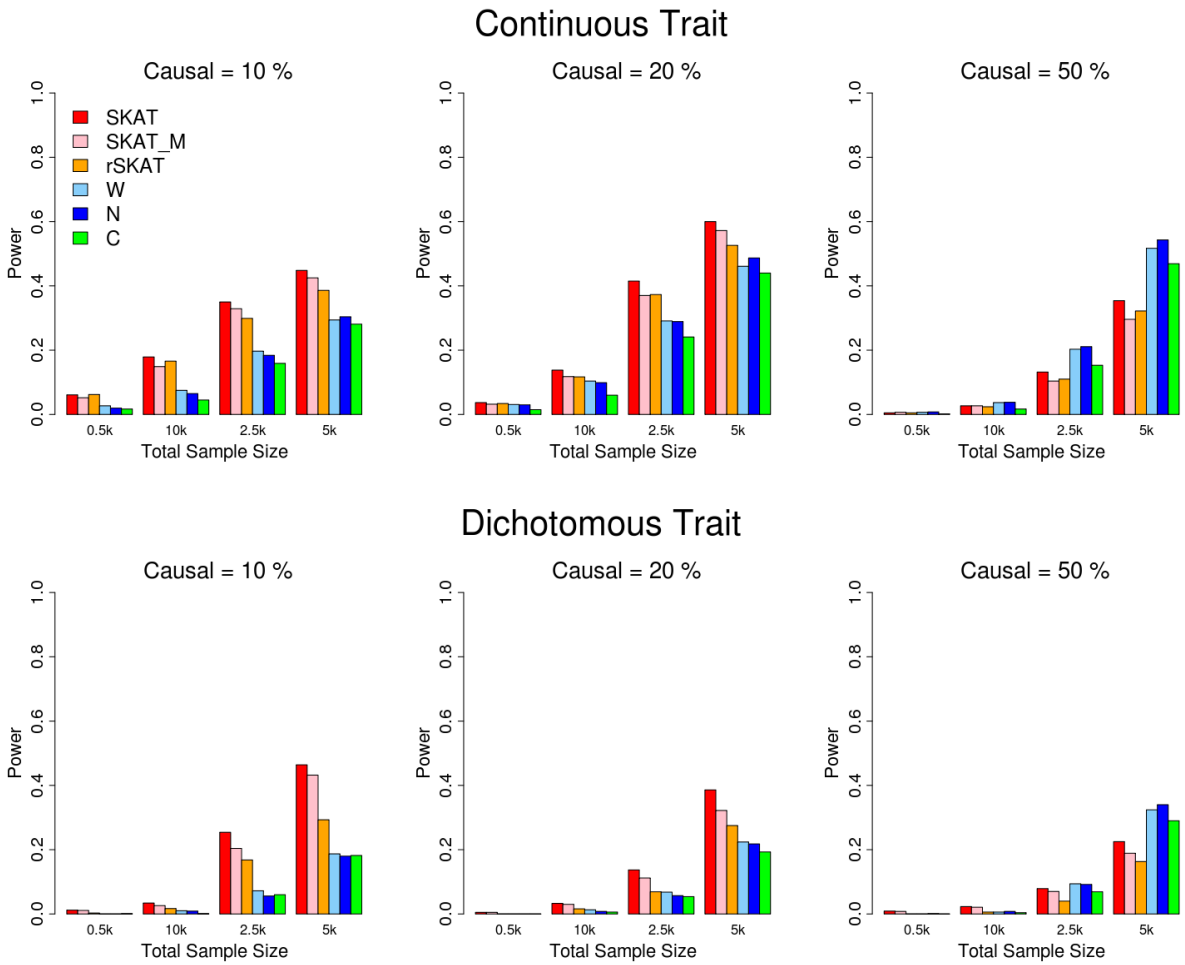

Figure S9. Power Comparisons for SKAT and Burden Tests Using Simulation Studies When Region Size = 3 kb and β+/- = 100/0

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 100% positive/0% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

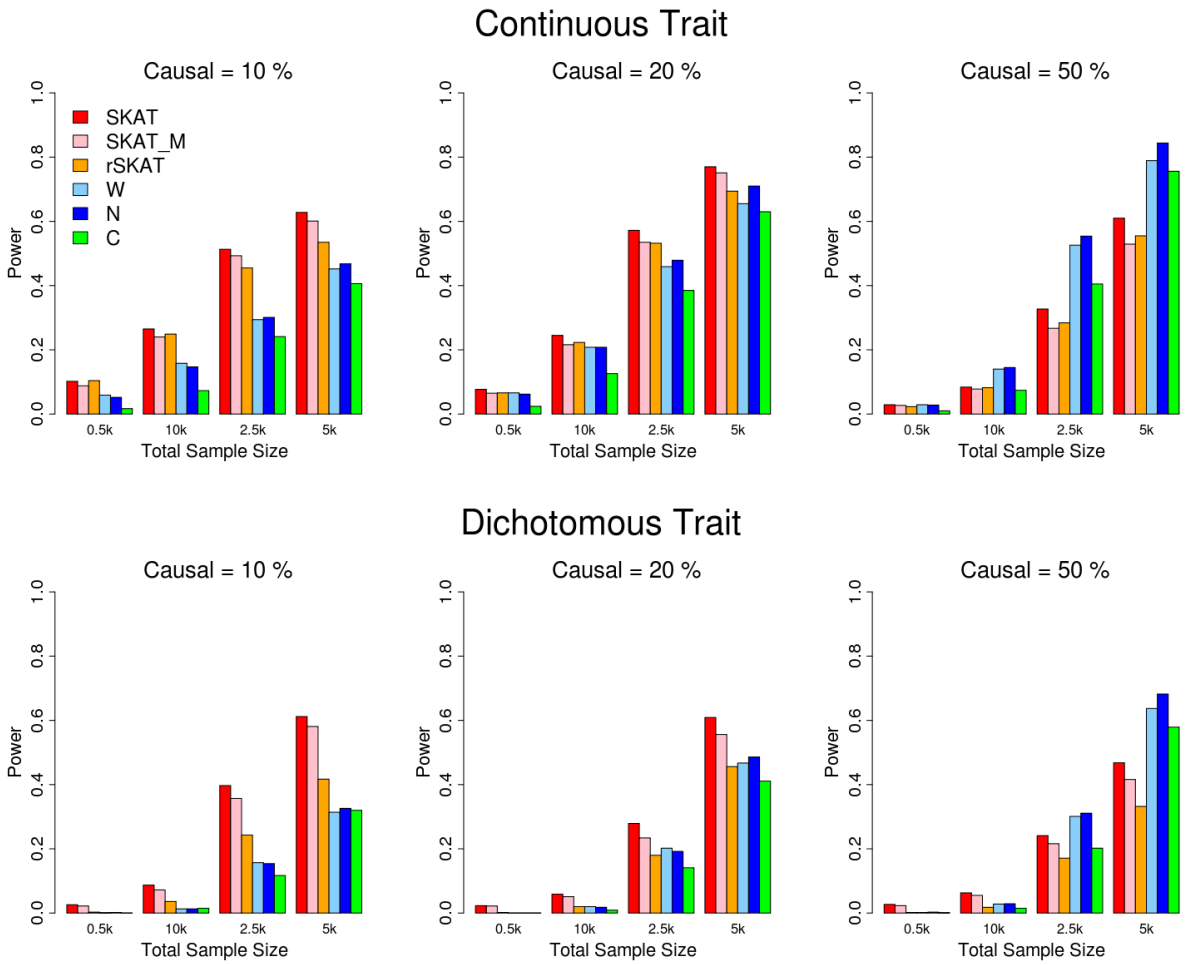

Figure S10. Power Comparisons for SKAT and Burden Tests Using Simulation Studies When Region Size = 5 kb and β+/- = 100/0

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 100% positive/0% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

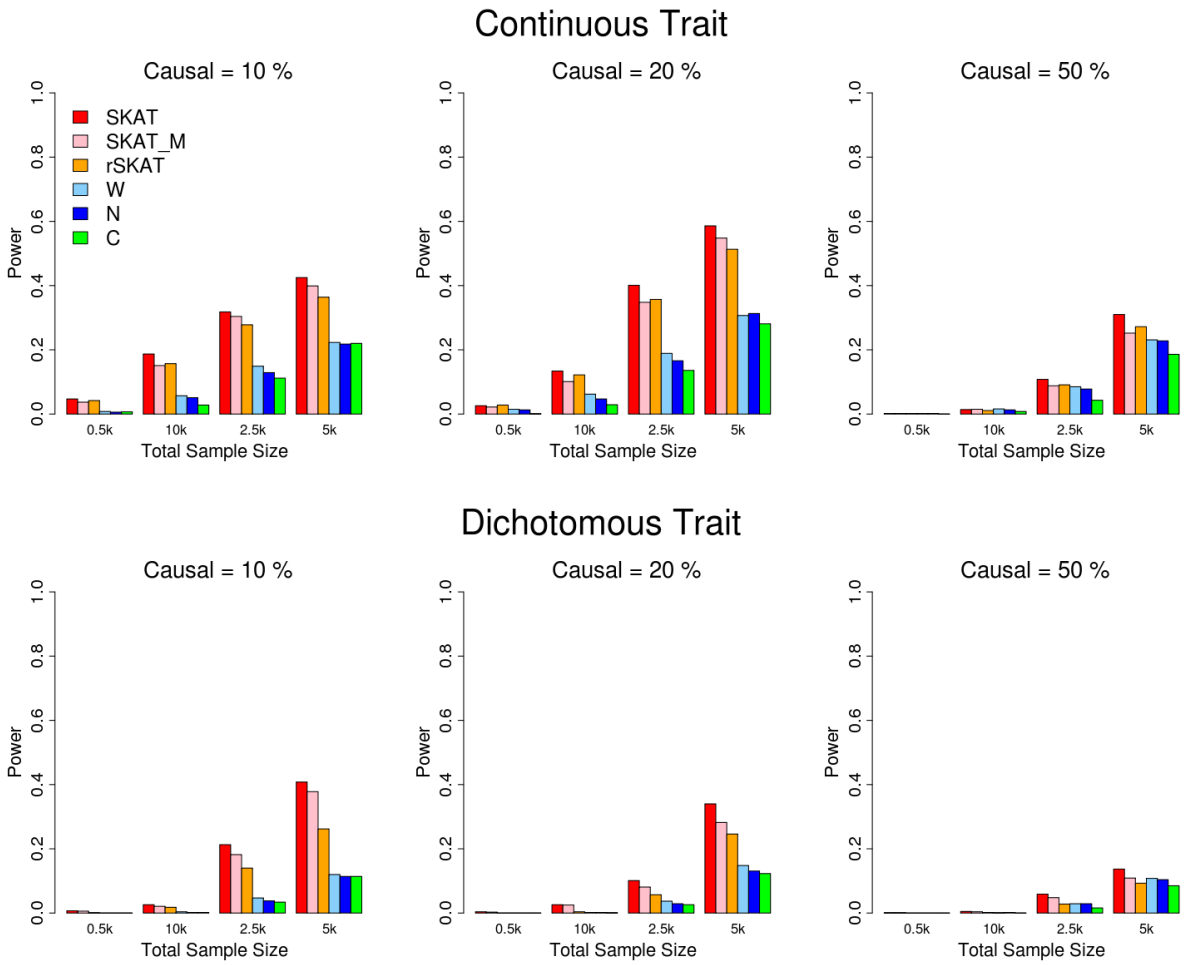

Figure S11. Power Comparisons for SKAT and Burden Tests Using Simulation Studies When Region Size = 3 kb and β+/- = 80/20

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 80% positive/20% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

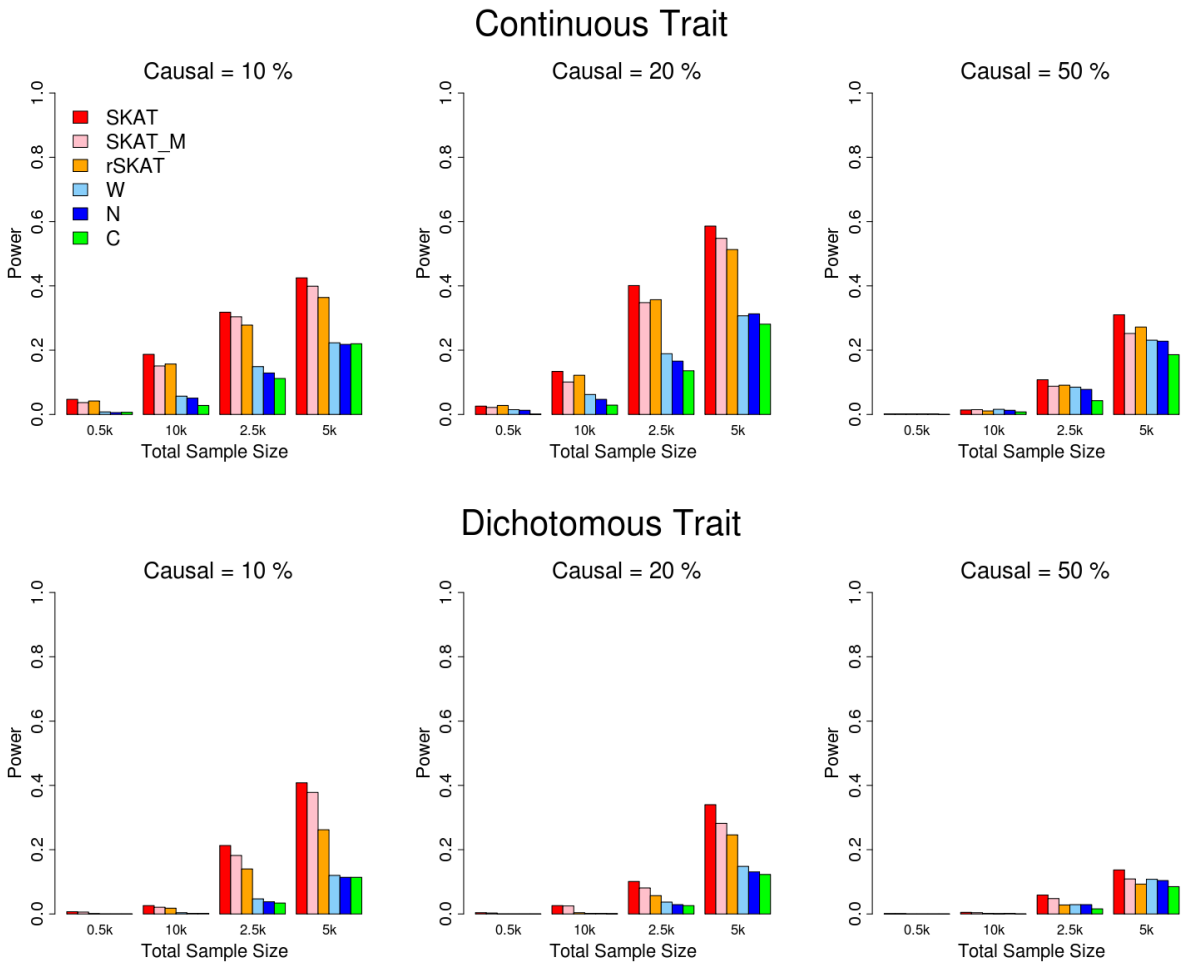

Figure S12. Power Comparisons for SKAT and Burden Tests Using Simulation Studies When Region Size = 5 kb and β+/- = 80/20

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 80% positive/20% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

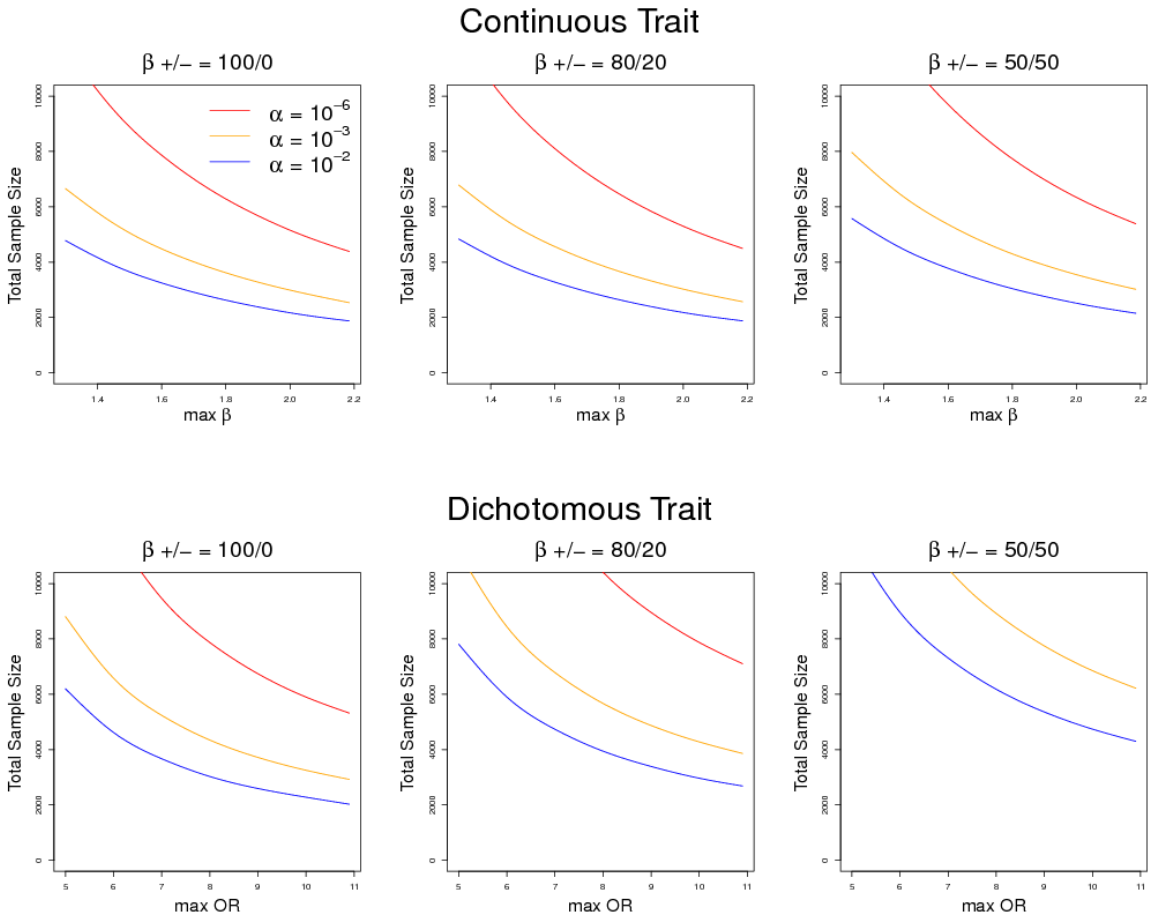

Figure S13. Required Sample Sizes to Reach 80% Power at the Genome-wide Level α=10^{-6} When Causal Variants Have MAF<1%

Required sample sizes to reach 80% power to detect rare variants associated with a continuous phenotype (top panel) or a dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the α=10^{-6} (red), 10^{-3} (orange) and 10^{-2} (blue) level as estimated by the analytical formulae using beta(0.5 , 0.5) weights under the same setting as Figure 1 except that the 5% of variants with MAF<1% are generated as causal variants. Required sample sizes are plotted against the "maximum" effect sizes (ORs) when MAF=10^{-4}. Estimated total sample sizes were averaged over 100 random 30kb regions. See **Methods and Supplementary Materials.**

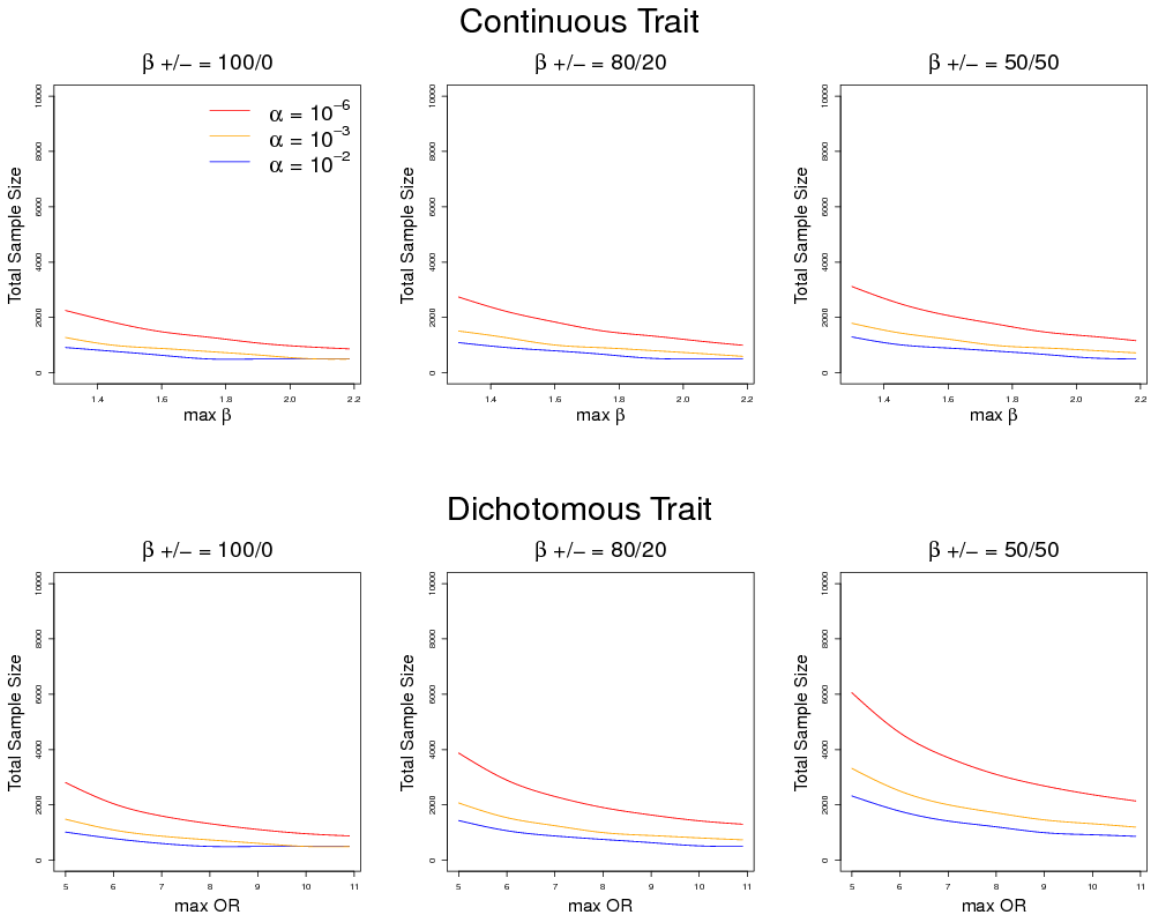

Figure S14. Required Sample Sizes to Reach 80% Power at the Genome-wide Level $\alpha = 10^{-6}$ When 10% of Rare Variants Are Causal Variants

Required sample sizes estimated analytically to reach 80% power to detect rare variants associated with a continuous (top panel) or dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the $\alpha=10^{-6}$, 10^{-3}, and 10^{-2} level. The setting is the same as Figure 2 except that a 10% of rare variants with MAF<3% within the 30 kb regions are causal. Plots correspond to 100%, 80%, and 50% of the causal variants associated with increase in the continuous phenotype or risk of the dichotomous phenotype. Regression coefficients for the *s* causal variants were assumed to be the same decreasing function of MAF as that in Figure 1. Required total sample sizes are plotted again the "maximum" effect sizes (ORs) when MAF=10^{-4}. Estimated total sample sizes were averaged over 100 random 30kb regions.

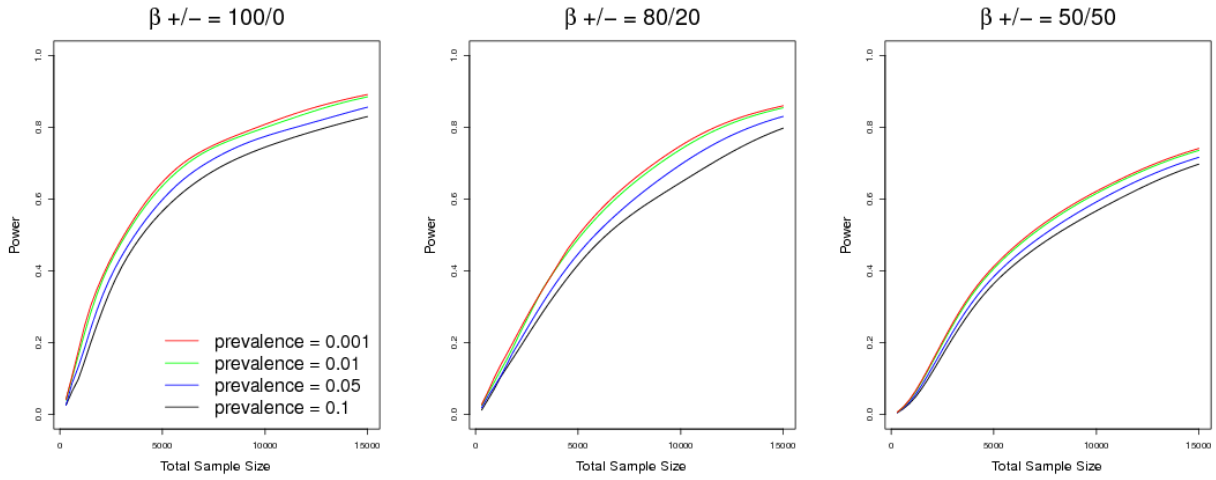

Figure S15. Estimated Power to Detect Rare Variants Association with a Dichotomous Case-Control Phenotype for Different Prevalences

Power to detect rare variants association with a dichotomous case-control phenotype estimated by the analytical formulae under the same setting as Figure 2 except that the disease penetrance = 0.1 (black), 0.05 (blue), 0.01 (green) and 0.001 (red). Plots correspond to the scenarios when 100%, 80%, and 50% of the causal variants are associated with increase in the risk of the dichotomous phenotype. Estimated powers were averaged over 100 random 30kb regions.