The American Journal of Human Genetics Volume *87*

Supplemental Data

Extending Rare-Variant Testing Strategies:

Analysis of Noncoding Sequence and Imputed Genotypes

Matthew Zawistowski, Shyam Gopalakrishnan, Jun Ding, Yun Li, Sara Grimm, and Sebastian Zöllner

Figure S1. Type I Error Rates for the Rare Variant Tests. Power is computed at α=0.01 on datasets with N=1000 cases and controls at the indicated neutral inclusion rates (p_n). Type I error for the CMAT, WSS and collapsing (COLL) tests is well-controlled across values of p_n. Type I error for the private allele test (PRIV) is initially conservative then increases with p_n becoming anti-conservative.

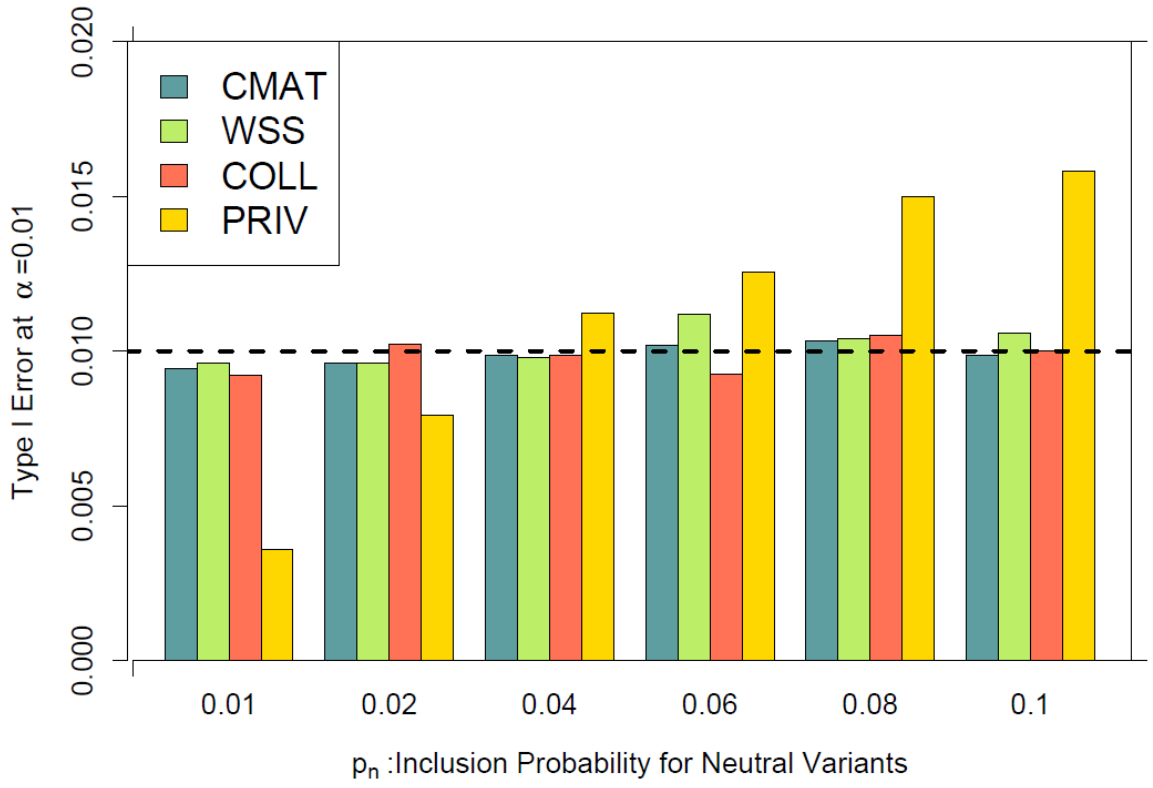

Figure S2. Power for CMAT and WSS using both uniform weighting (Eq. 2) and maf-based weights described in [13]. Conditional on weighting scheme, the CMAT and WSS have similar power across the grid. The maf-based weights correspond more closely to our disease model (Fig. 1) than do the simple uniform weights and therefore provided a more powerful analysis for both methods except when misspecification rates are highest.

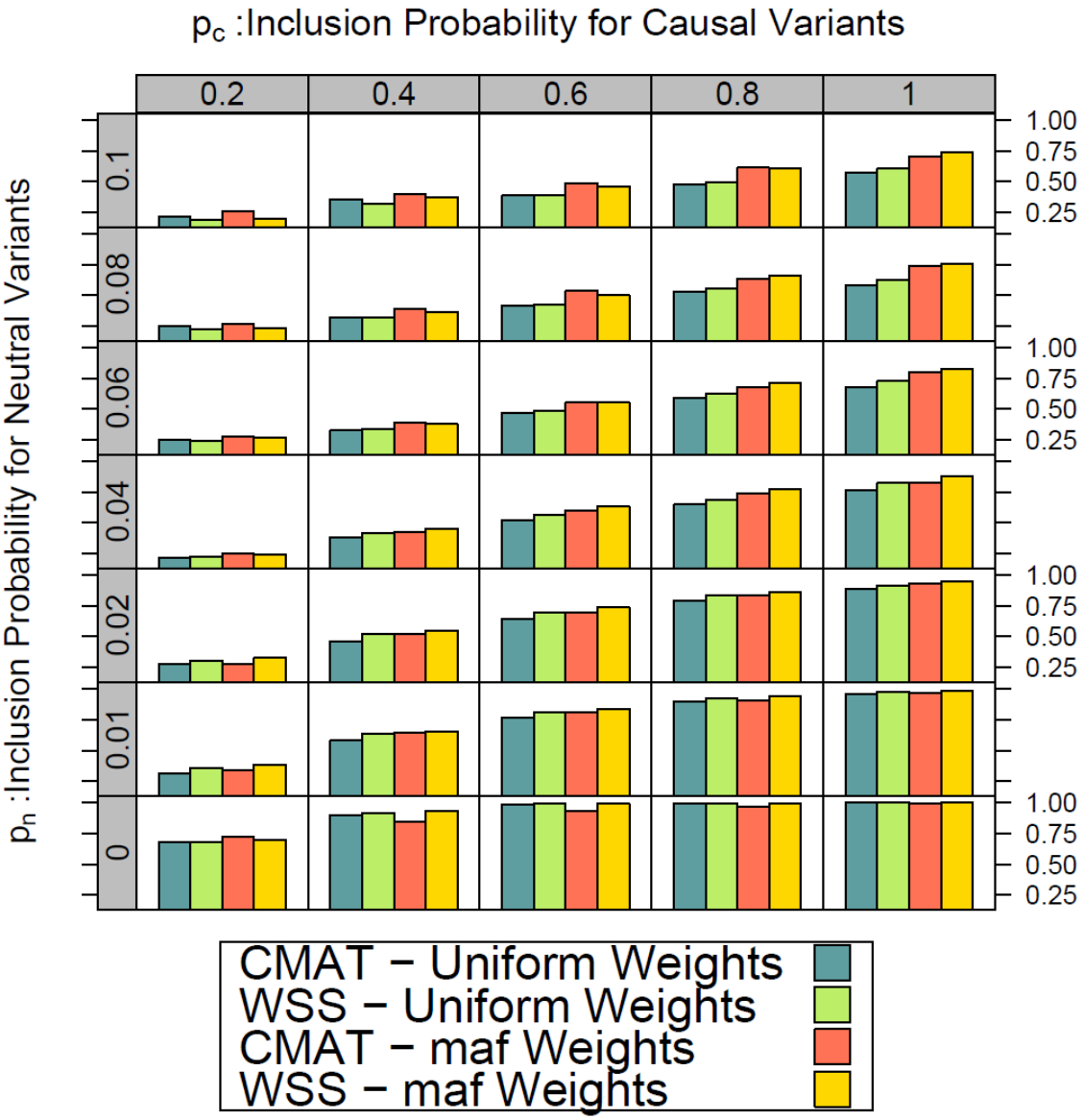

Figure S3. Power to analyze Deep Sequence Datasets for minor allele cutoff of β=1%. Each dataset contains exact genotypes for N=1000 cases and controls based on k=15 causative variants in the population. Along the vertical axis we vary the probability of (incorrectly) including a neutral variant (pn) in the analysis and along the horizontal axis we vary the probability of (correctly) including a causative variant (pc). The height of the bars in each cell indicates the power for the four tests at α=0.01. Here, the maximum allele frequency for risk variants (p_{max}) was set to 1%.

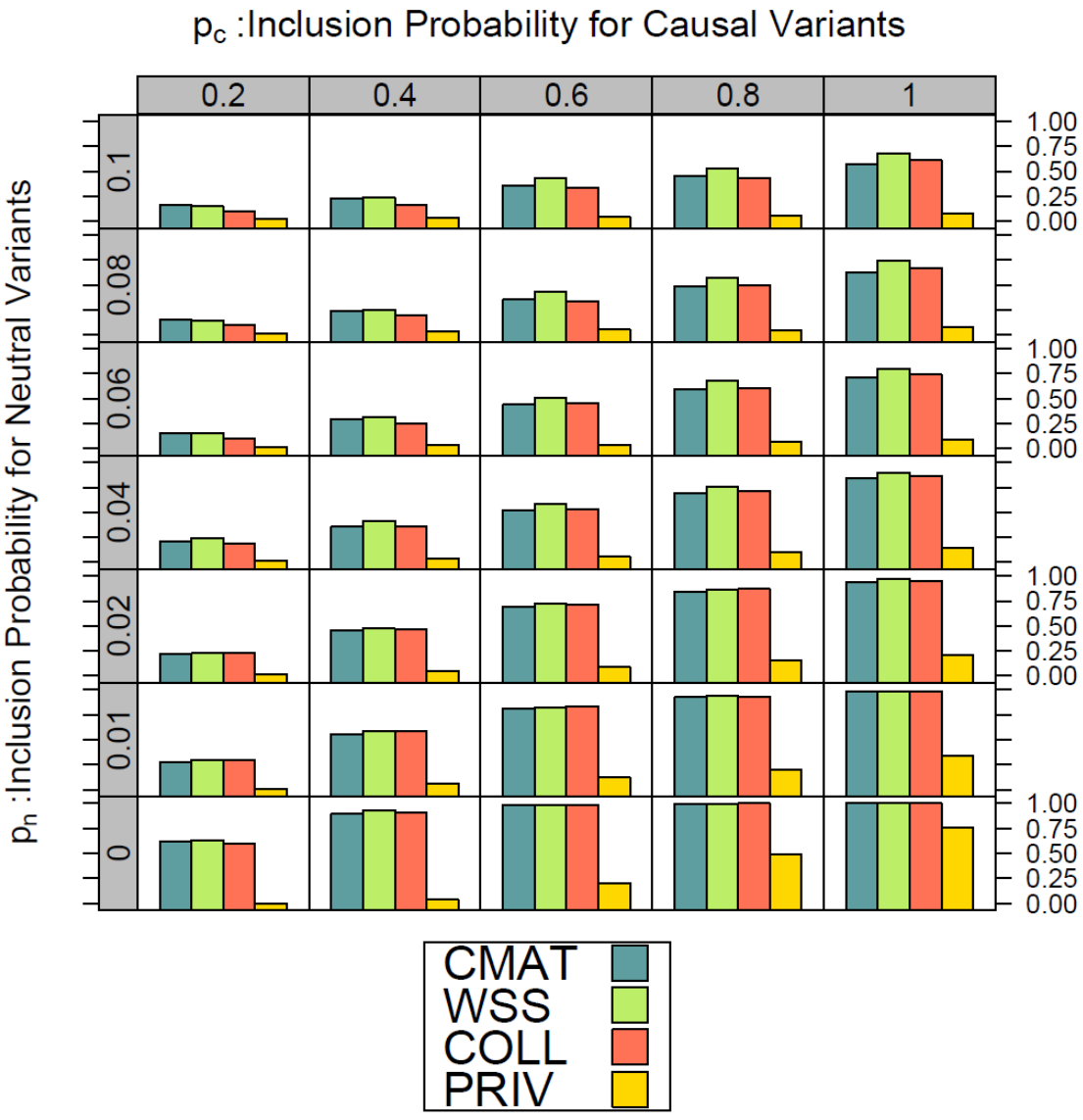

Figure S4. CMAT power for Imputation Datasets with minor allele cutoff of β=1%. Datasets contain exact genotypes for N_{seq} cases and controls and probabilistic genotypes based on imputation for the remaining samples. The top line shows CMAT power for Deep Sequencing datasets (N_{seq} = N) and serves as an upper bound for power at a fixed total sample size N. Here, the maximum allele frequency for risk variants (p_{max}) was set to 1%. We report power at α=0.01 using the whole-gene inclusion threshold (p_n = 0.1; p_c = 0.8).