Supplementary Materials for “Ordered Multinomial Regression for Genetic Association Analysis of Ordinal Phenotype at Biobank Scale”

Christopher A. German[1] Janet S. Sinsheimer[1,2,3] Yann Klimentidis[4]
Hua Zhou[1] Jin J. Zhou[4]

[1]Department of Biostatistics, UCLA Fielding School of Public Health
[2]Department of Human Genetics, David Geffen School of Medicine at UCLA
[3]Department of Computational Medicine, David Geffen School of Medicine at UCLA
[4]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona

S.1 Summary Statistics of Non-Hispanic White (NHW) COPDGene Cohort Data

Summary statistics of the COPDGene cohort used in the analysis are shown in Table 1. There were 19 individuals with missing data and 698 individuals with unidentifiable GOLD category (-1: FEV1/FVC ratio was above 0.7, but percent predicted FEV1 was less than 80%).

For the genotype data, histograms of the MAF values, number of missing SNPs per person, and number of missing people per SNP are shown in Figure 1. The mean MAF is 0.245.

S.2 Additional Information on NHW COPDGene Results

LocusZoom plots of the top hits from the COPDGene data using the ordered multinomial method are shown in Figure 2. In the plot on Chromosome 3, the top hit is nearest to the EEFSEC gene, shown by Hobbs et al. (2017) to be associated with COPD.

Decreased power resulted from analysis (logistic regression) of extreme values (analyzing only those labeled *not a case* and *GOLD stage 4*). Manhattan plot for this analysis is shown in Figure 3.

Linear regression GWAS on the ordinal hypertension trait and the Manhattan plot for these results is shown in Figure 4.

S.3 Summary Statistics of UK Biobank data

Summary statistics of the UK Biobank phenotype data used in the analysis are shown in Table 2. Individuals who met the genotype filtering and missingness criteria were not on medication.

S.4 Additional Analysis Information of UK Biobank data

Information on the top hits ($P \leq 10^{-8}$) from OrdinalGWAS on the UK Biobank data is included in the Supplementary Table in Table 3. It contains locus information where applicable. QQ plots from this analysis are shown in Figure 5.

We ran a sex by SNP GWAS to demonstrate G×E capability of our software on these top hits. The results from this analysis is in the Supplementary Table in Table 4.

To demonstrate the capability of our software to run SNP-set analysis, we performed SNP-set GWAS on the UK Biobank data with a window size of 20 SNPs. The Manhattan plot from this analysis is shown in Figure 6.

S.5 Additional Simulation Results on SNP-set and G×E

We looked at the power of the methods in a SNP-set setting. We used the 11 SNPs from the COPDGene data that are on the CHRNA5 gene on chromosome 15 that was found to be associated with COPD ($n = 6670$). We then ran 1000 replicates generating a trait under different effect sizes where 3 of the 11 SNP were randomly selected to be causal. We used four categories with the threshold value $\boldsymbol{\theta} = (0.1, 3.0, 3.1)$. We used the same link function and covariates from the original power analysis, with the effect sizes set to be 1.0 and 2.0 for standardized age and sex respectively. We generated the trait according to the proportional odds model. The power plot for this analysis is shown in Figure 7.

We also investigated the power of the method in a G×E setting. We used four categories with the threshold value $\boldsymbol{\theta} = (0.1, 3.0, 3.1)$. The same covariates, covariate effect sizes, and link function as the original power analysis were used. The SNP was generated with a minor allele frequency of 0.2 according to Hardy-Weinberg equilibrium and the SNP effect size was fixed at 0.05. For the environmental variable for the G×E effect, the standardized age variable was used and the effect size of the interaction was varied from 0.0 to 0.5 in increments of 0.05. The power plot for this analysis from 1000 replicates is shown in Figure 8.

S.6 Comparison of Score, Wald, and Likelihood Ratio Tests

To assess the power of score test in a GWAS setting, we compared the Wald test, LRT, and score test under the same setting as in the simulations. We used a MAF of 0.20, $\boldsymbol{\theta}$ = (0.1, 3.0, 3.1), and checked varying sample sizes. We ran the three tests at different effect sizes γ. We completed 1000 replicates for two sample sizes shown in Figure 9 and looked at power at the 10^{-6} significance level. Even at a sample size of 500, the score test did nearly the same in terms of power as the likelihood ratio test. At $n = 2500$, all three tests had equal power. This provides the justification for using a score test in a GWAS setting, especially at biobank scale.

Table 1: Summary statistics for Non-Hispanic White individuals used in analysis from COPDGene cohort. $^{\#}$: n (%), $^{\$}$: Mean (SD).

Variable	Value
n	6678
Gold$^{\#}$	
0	2534 (37.9)
1	607 (9.1)
2	1424 (21.3)
3	909 (13.6)
4	487 (7.3)
-1	698 (10.3)
NA	19 (0.5)
Gender$^{\#}$	
Male	3180 (47.6)
Age Enroll$^{\$}$	62.1 (8.8)
ATS Pack Years$^{\$}$	47.3 (26.0)
FEV1$^{\$}$	2.2 (1.1)
FVC$^{\$}$	3.3 (1.2)
FEV1 FVC ratio$^{\$}$	0.6 (0.5)
FEV1 Percent Predicted$^{\$}$	73.3 (26.3)
Height (cm)$^{\$}$	169.7 (9.5)

Table 2: UK Biobank baseline characteristics. $^{\#}$: n (%), $^{\$}$: Mean (SD).

Variable	Value
n	185565
Hypertension$^{\#}$	
0	30822 (16.6)
1	25242 (13.6)
2	50461 (27.2)
3	75286 (40.6)
4	3754 (2.0)
Gender$^{\#}$	
Male	85125 (45.9)
Age$^{\$}$	56.5 (8.1)
Average SBP$^{\$}$	137.2 (18.5)
Average DBP$^{\$}$	81.8 (10.1)
BMI$^{\$}$	27.1 (4.6)

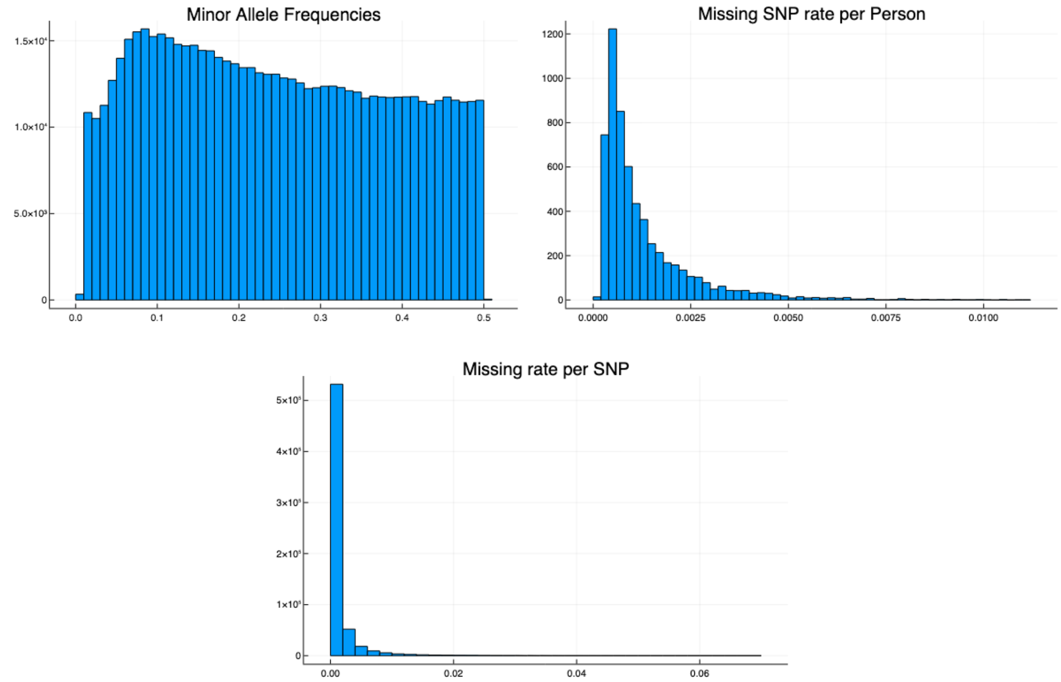

Figure 1: Histograms of minor allele frequencies, missing SNPs per person, and missing people per SNP of NHW COPDGene genotype data.

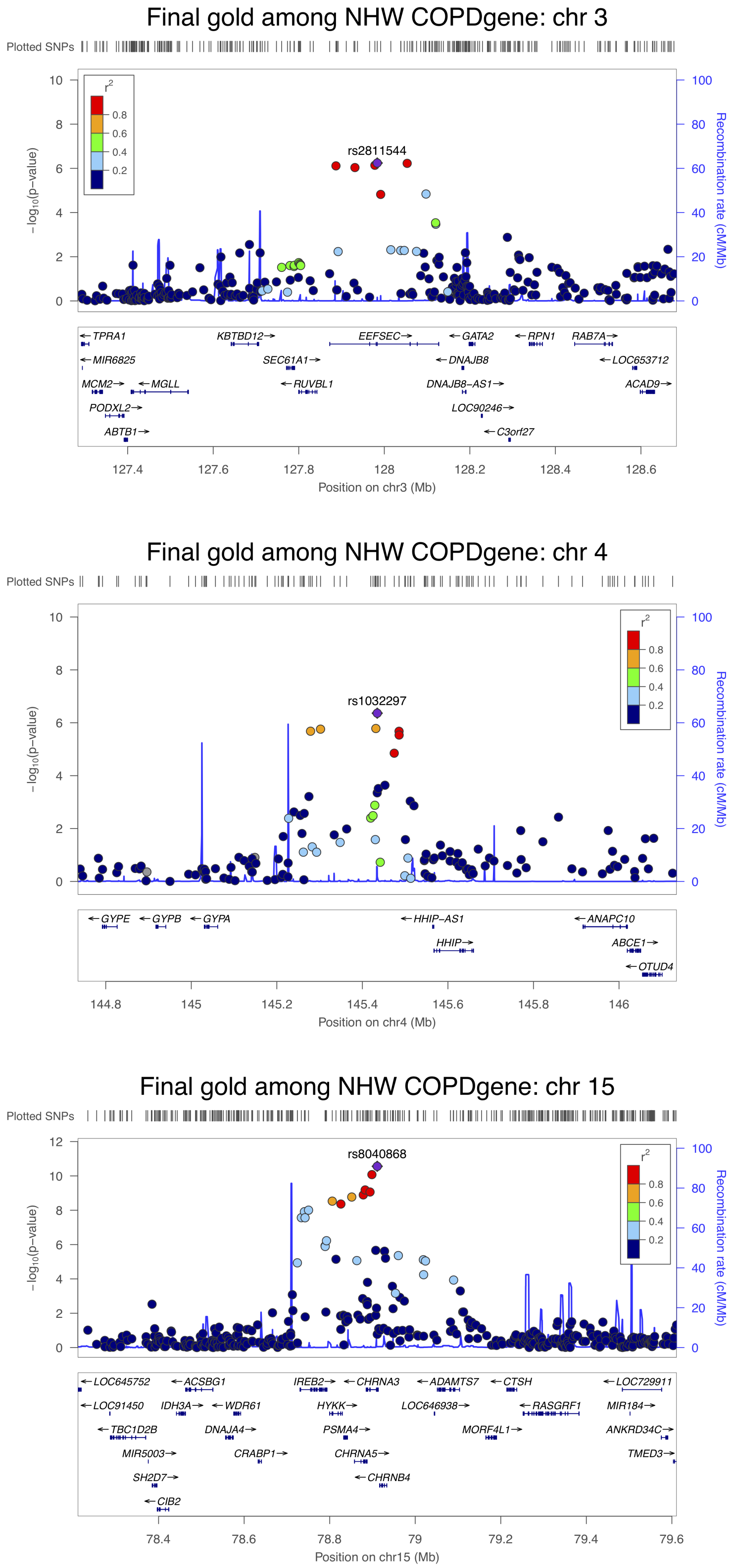

Figure 2: LocusZoom Plots, linkage disequilibrium and recombination rates around the genome-wide significant top hits using the ordered multinomial regression on the COPDGene data.

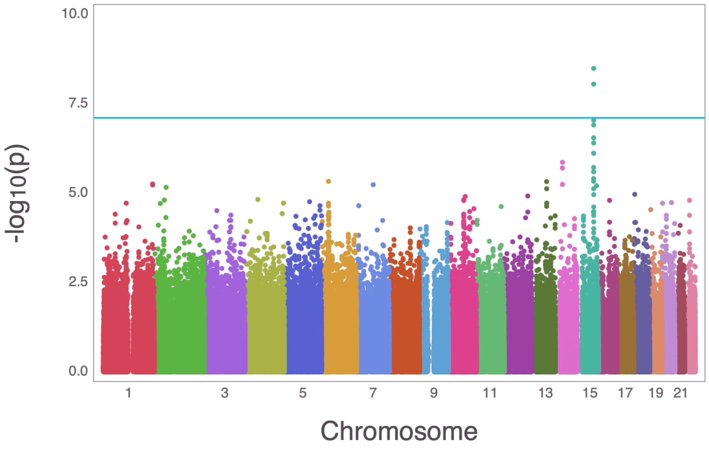

Figure 3: Manhattan plot on COPDGene analysis (logistic regression) analyzing only those labeled *not a case* and *GOLD stage 4*.

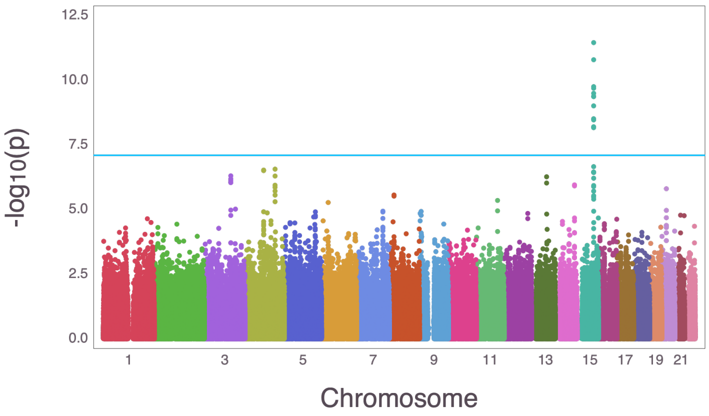

Figure 4: Manhattan plot on COPDGene analysis (linear regression) run on the ordinal GOLD stage variable.

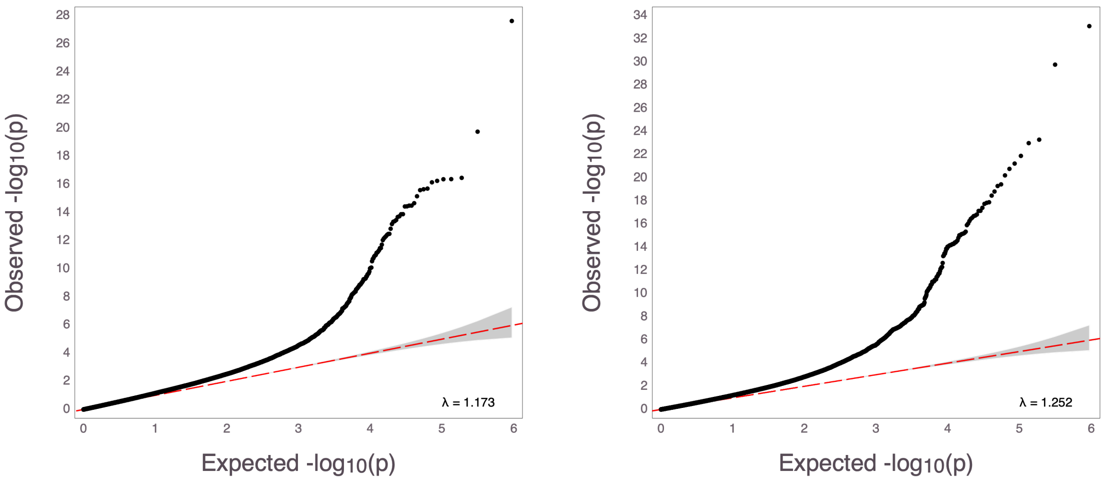

Figure 5: QQ plots for the hypertension GWAS results in UK Biobank. Left is the QQ plot using logistic regression. Right is the QQ plot using ordered multinomial regression. λ is the genomic inflation factor.

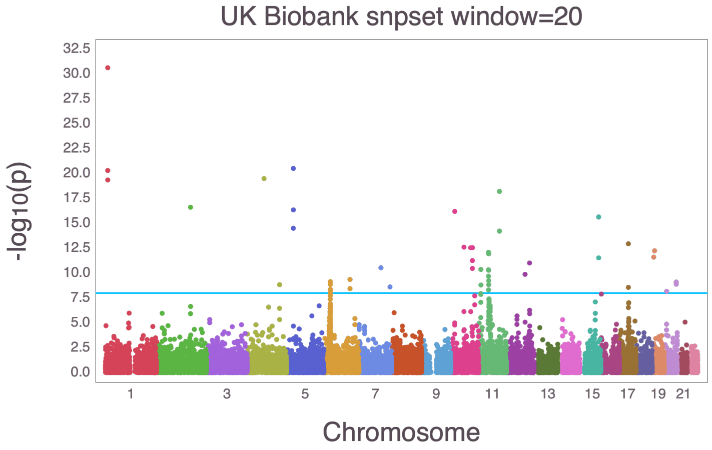

Figure 6: Manhattan plot of UK Biobank SNP-set analysis (every 20 snps) using OrdinalGWAS.

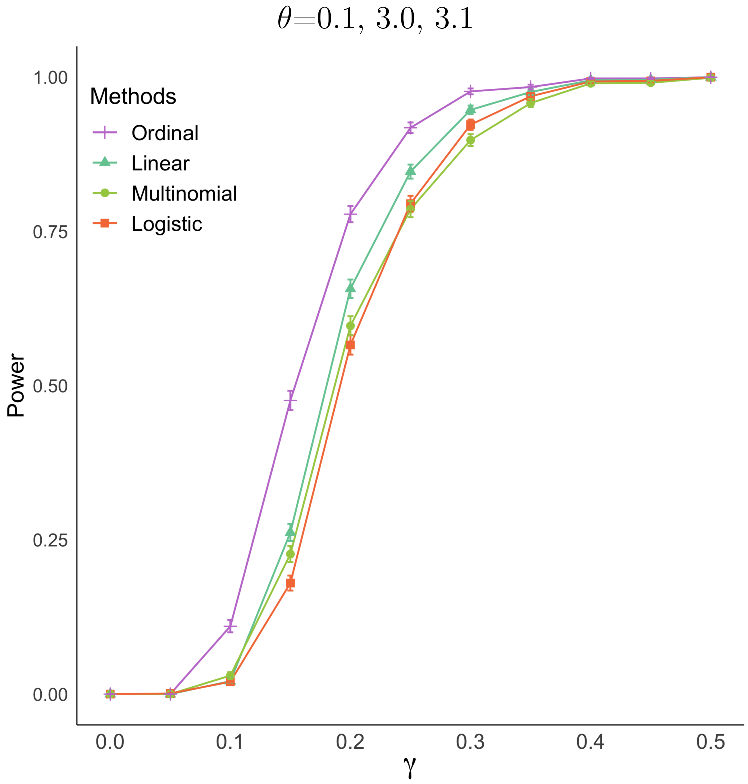

Figure 7: Power comparisons of SNP-set analysis from 1000 simulation replicates. One SNP-set with 11 SNPs from CHRNA5 of COPD data were used. Trait were simulated under different effect sizes where 3 of the 11 SNPs were causal with the same contribution.

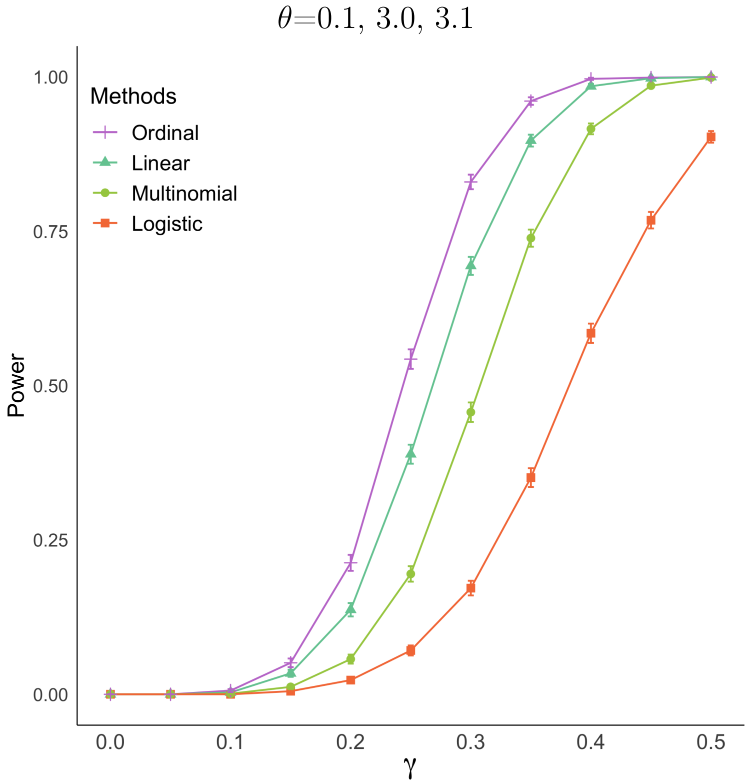

Figure 8: Power comparisons for G×E analysis using 1000 simulation replicates. Four categories of $\boldsymbol{\theta} = (0.1, 3.0, 3.1)$ were used. The SNP was generated with a minor allele frequency of 0.2 according to Hardy-Weinberg equilibrium and the SNP effect size was fixed at 0.05. For the G×E effect, the standardized age variable was used and the effect size of the interaction was varied from 0.0 to 0.5 in increments of 0.05. The same covariates, covariate effect sizes, and link function as the original power analysis were used in the main text.

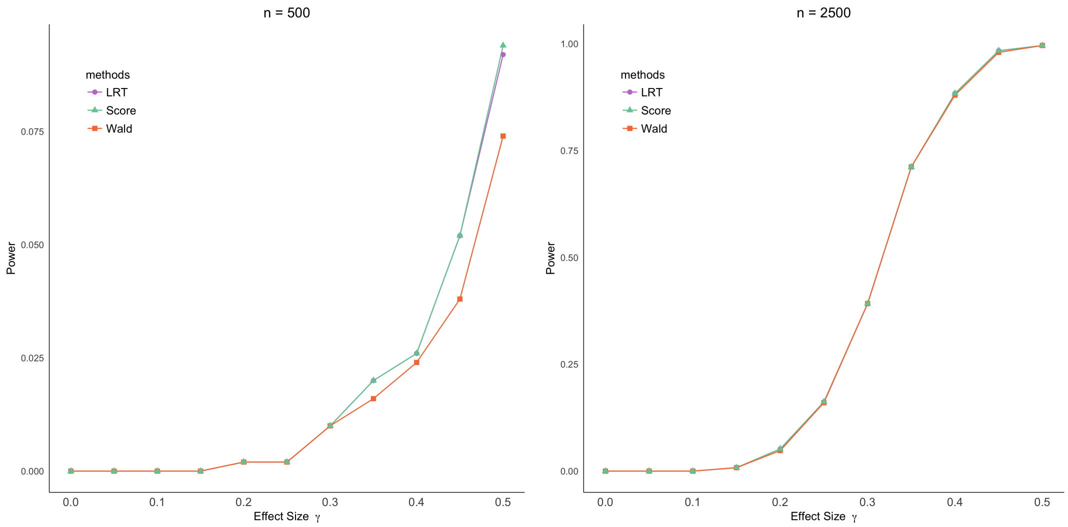

Figure 9: Comparison of power between using a likelihood ratio test (LRT), score test, and a Wald test. The left plot is for a sample size of 500, the right plot is for replicates with a sample size of 2500.