Supplementary material for "Genome-wide association test of multiple continuous traits using imputed SNPs"

Baolin Wu[1], James S. Pankow[2]

[1]Division of Biostatistics, [2]Division of Epidemiology and Community Health

School of Public Health, University of Minnesota

1 Simulation with rare variants

Here we conducted simulations to investigate the performance of various methods for testing less frequent and rare variants. We follow the same simulation setup as described in the main paper, and report the results for the following scenarios: (1) $n = 2000$, $p_0 = 0.1, p_1 = 0.1, \rho = 0.5$, $\tau = 0.95$; (2) $n = 5000$, $p_0 = 0.05, p_1 = 0.05, \rho = 0.5$, $\tau = 0.95$; (3) $n = 5000$, $p_0 = 0.01, p_1 = 0.01, \rho = 0.5$, $\tau = 0.95$. Here we have assumed larger sample sizes for rare variants.

Table 1 summarizes the estimated type I errors. When estimating parameters for the POM GEE tests, we have used the iterative re-weighted least squares algorithm, which had convergence problems leading to missing results for rare variants in the null simulations. For the ACL based tests, using the "best-guess" genotypes leads to slightly conservative type I errors compared to their corresponding ACL GEE tests. All tests control the type

I error rate reasonably well. The ACL GEE tests generally have better performance for relatively common variants and have subpar performance for rare variants. While both MLM GEE tests and POM test using "best-guess" genotypes consistently perform well under both relatively common and rare MAFs.

Table 1: Type I error of testing four continuous traits. The MAFs of SNP are p_0 and $p_0 + p_1$ in the two populations. We set $p_1 = p_0$ and $\tau = 0.95$. Q is the 4-DF omnibus test, T and T' are the 1-DF tests assuming common or common scaled effect. (Q_a, T_a, T'_a) are the ACL GEE tests. (Q_o, T_o, T'_o) are the POM GEE tests. (Q_s, T_s, T'_s) are the MLM GEE tests. $(\tilde{Q}_a, \tilde{T}_a, \tilde{T}'_a)$ are the ACL tests using the "best-guess" genotypes. $(\tilde{Q}_o, \tilde{T}_o, \tilde{T}'_o)$ are the POM tests using the "best-guess" genotypes. The type I errors have been scaled by the nominal significance level α.

							$p_0 = 0.10$								
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.19	1.04	1.10	0.95	0.84	0.90	NA			0.89	0.86	0.87	1.02	0.88	0.99
10^{-3}	1.12	1.08	1.08	0.95	0.94	0.90	NA			1.02	0.96	0.98	0.97	0.99	1.00
10^{-2}	1.05	1.03	1.03	0.95	0.97	0.97	NA			0.99	0.99	0.98	0.97	1.00	1.00
							$p_0 = 0.05$								
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.18	1.09	1.08	0.88	0.83	0.85	NA			0.90	0.77	0.79	1.13	0.89	0.96
10^{-3}	1.09	1.03	1.01	0.97	0.97	0.94	NA			0.98	0.95	0.95	0.97	0.97	0.93
10^{-2}	1.04	1.01	1.01	0.98	0.98	0.96	NA			0.96	0.98	0.97	0.97	0.98	0.98
							$p_0 = 0.01$								
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.30	1.09	1.25	0.76	0.96	0.95	NA			0.92	1.02	0.99	0.67	0.83	0.86
10^{-3}	1.17	1.13	1.16	0.90	0.94	0.92	NA			0.90	0.96	0.95	0.78	0.88	0.90
10^{-2}	1.13	1.10	1.08	0.95	0.98	0.97	NA			0.96	0.97	0.96	0.95	0.97	0.97

Table 2 summarizes the power under significance level $\alpha = 10^{-4}$. Here we have assumed larger effect sizes for rare variants (MAF=0.01). When estimating parameters for the POM GEE tests, we have used the iterative re-weighted least squares algorithm, which had convergence problems leading to missing results for rare variants in the simulations. The 1-DF tests are the most powerful when either γ_j or γ_j/σ_j are close to each other. ACL tests using the "best-guess" genotypes had reduced power compared to the ACL GEE tests. The ACL GEE tests have comparable performance as the MLM GEE tests. Interestingly the POM tests using the "best-guess" genotypes have comparable power as

Table 2: Power of testing four traits at significance level $\alpha = 10^{-4}$. The MAFs of SNP are p_0 and $p_0 + p_1$ in the two populations. We set $p_1 = p_0$ and the SNP imputation uncertainty parameter $\tau = 0.95$. γ_i is the SNP coefficient.

$p_0 = 0.10$

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.775	0	0.004	0.719	0	0.004		NA		0.770	0	0.005	0.766	0	0.004
(0.3,0.2,0.1,0)	0.879	0.041	0.220	0.839	0.031	0.188		NA		0.878	0.039	0.219	0.873	0.051	0.222
(.25,.18,.18,.18)	0.486	0.658	0.725	0.422	0.599	0.667		NA		0.473	0.649	0.714	0.469	0.651	0.717
(0.2,0.2,0.2,0.2)	0.622	0.831	0.773	0.556	0.784	0.718		NA		0.609	0.823	0.760	0.606	0.827	0.769

$p_0 = 0.05$

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.876	0	0.006	0.803	0	0.005		NA		0.909	0	0.007	0.872	0	0.006
(0.3,0.2,0.1,0)	0.946	0.063	0.307	0.901	0.051	0.257		NA		0.966	0.075	0.356	0.944	0.073	0.311
(.25,.18,.18,.18)	0.615	0.778	0.832	0.517	0.690	0.755		NA		0.673	0.821	0.868	0.605	0.772	0.829
(0.2,0.2,0.2,0.2)	0.751	0.910	0.869	0.660	0.851	0.803		NA		0.806	0.937	0.903	0.744	0.910	0.867

$p_0 = 0.01$

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)/2$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.155	0	0.002	0.092	0	0.001		NA		0.353	0	0.002	0.138	0	0.001
(0.3,0.2,0.1,0)	0.224	0.008	0.032	0.148	0.005	0.022		NA		0.481	0.015	0.075	0.205	0.009	0.030
(.25,.18,.18,.18)	0.067	0.142	0.171	0.034	0.090	0.110		NA		0.157	0.294	0.343	0.058	0.132	0.161
(0.2,0.2,0.2,0.2)	0.103	0.241	0.199	0.056	0.159	0.128		NA		0.240	0.452	0.393	0.089	0.224	0.190

the ACL and MLM GEE tests for relatively common variants ($p_0 = 0.1$), and perform better for rare variants ($p_0 = 0.05, 0.01$).

2 Selection of genetic model

In practice the additive model has been the most widely used model in testing genetic associations. Another two genetic models, recessive or dominant, can also be applied. For these two alternative models, we can collapse the three genotypes into two groups correspondingly and model the collapsed genotypes with a Bernoulli distribution. We can then similarly derive a conditional (on covariates and outcomes) logistic regression model for the collapsed genotypes, which can be used to test the multi-trait association. Our previously derived fractional multinomial model can be applied to any K-category multinomial distribution, and hence the same model and GEE estimation can be applied to model the collapsed genotypes.

Specifically for the recessive model, we study the following fractional logistic regression model based quasi-likelihood, $(p_{i0}+p_{i1})\log(\phi_{i0}+\phi_{i1})+p_{i2}\log(\phi_{i2})$, where $\log\frac{\phi_{i2}}{\phi_{i0}+\phi_{i1}}$ is modeled as a linear function of covariates and outcomes; and for the dominant model, we study the following fractional logistic regression model based quasi-likelihood, $p_{i0}\log(\phi_{i0}) + (p_{i1}+p_{i2})\log(\phi_{i1}+\phi_{i2})$, where $\log\frac{\phi_{i2}+p_{i1}}{\phi_{i0}}$ is modeled as a linear function of covariates and outcomes. Here (p_{i0}, p_{i1}, p_{i2}) are the imputation scores for individual i, and $\phi_{ik} = \Pr(G_i = k|X_i, Y_i)$, where G_i is the genotype score, and X_i and Y_i are the covariate and outcome vectors respectively.

3 Joint test of mixed outcomes

Strictly speaking, only for normally distributed multiple continuous traits under multivariate linear regression models, we can derive the ACL model for the conditional genotype distributions, whereas the POM model closely approximates the ACL model (Wu and Pankow, 2015). For a set of mixed continuous and discrete outcomes, in general it is very hard to jointly model their distributions and correspondingly derive an inverted regression model analytically. A simple approach is to include all outcomes in the inverted regression model. The MLM GEE test approach of He *et al.* (2013) can be generally applied. Here we conduct a simple simulation study to briefly compare the performance of ACL/POM models and the MLM GEE test.

We consider a mix of one continuous outcome Y and one binary outcome D, and simulate two covariates (X_1, X_2) and the genotype G as previously. We simulate the outcome dependence following the approach of Ghosh *et al.* (2013). Specifically we simulate a zero-mean bivariate normal vector (ϵ_1, ϵ_2) with variance $(2, 1)$ and correlation ρ. We then transform the normal error component ϵ_2 with

$$e_2 = \log\frac{\Phi(\epsilon_2)}{1-\Phi(\epsilon_2)},$$

Table 3: Power of testing two traits at significance level $\alpha = 10^{-5}$. The MAFs of SNP are 0.3 and 0.4 in the two populations. γ_i is the SNP coefficient.

(γ_1, γ_2)	ACL	POM	MLM GEE
(0.3,0)	0.397	0.401	0.406
(0.3,0.18)	0.410	0.413	0.423
(0.18,0.3)	0.073	0.073	0.074

where Φ is the standard normal cumulative distribution function. We can easily check that e_2 follows a logistic distribution. Define $Y = \mu_1 + \epsilon_1$ and $D = I(\mu_2 + e_2 > 0)$, where the two mean components μ_1 and μ_2 are linear functions of the covariates and genotype: $\mu_1 = 1 + 0.5X_1 + 0.5X_2 + \gamma_1 G$, $\mu_2 = 1 + X_1 + X_2 + \gamma_2 G$. Table 3 shows the power results for $\rho = 0.2$ at significance level $\alpha = 10^{-5}$ based on 10^4 simulations. Overall the MLM GEE test performs slightly better than the two inverted regression approaches. We have done simulations at other parameter settings, and have found that the naive approach of including all outcomes in the inverted regression model may not work as well as the MLM GEE test. It will be interesting to develop alternative inverted regression approach to jointly testing association of mixed outcomes, and further extend them to association test with imputed SNPs.

References

Ghosh,A., Wright,F.A. and Zou,F. (2013) Unified Analysis of Secondary Traits in Case-Control Association Studies. *Journal of the American Statistical Association,* **108** (502), 566–576.

He,Q., Avery,C.L. and Lin,D.Y. (2013) A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic Epidemiology,* **37** (8), 759–767.

Wu,B. and Pankow,J.S. (2015) Statistical methods for association tests of multiple continuous traits in genome-wide association studies. *Annals of human genetics,* **79** (4), 282–293.