Supplementary material for “Sequence kernel association test of multiple continuous phenotypes”

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 More simulation studies

As reviewer suggested, here we investigated the performance of different methods when assuming same direction of genetic effects. We use the same simulation setup as follows.

We simulated 1000 individuals and considered two covariates: a standard normal covariate X_1, and a binary ancestry indicator X_2 with $\Pr(X_2 = 1) = 0.5$. We consider testing $K = 4$ related traits with a compound-symmetry correlation matrix: $Y_1 = 1 + 0.5X_1 + 0.5X_2 + \eta_1 + \epsilon_1$, $Y_2 = 1 + X_1 + X_2 + \eta_2 + \epsilon_2$, $Y_3 = 1 + 0.5X_1 + 0.5X_2 + \eta_3 + \epsilon_3$, and $Y_4 = 1 + X_1 + X_2 + \eta_4 + \epsilon_4$, where $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ are zero-mean normal with variances $(\sigma_1^2 = 2, \sigma_2^2 = 1, \sigma_3^2 = 1, \sigma_4^2 = 1)$ and correlation ρ, and $(\eta_1, \eta_2, \eta_3, \eta_4)$ are contributions from the set of rare variants, which are simulated as follows.

Using a calibrated coalescent model (Schaffner *et al.*, 2005), we first generated 10,000 European-like haplotypes of length 1000 kb. Each time we randomly pair the haplotypes to simulate 1000 individuals. We study those rare variants with MAF ≤ 0.01 in a randomly selected gene region of length 10 kb, denoted as $(G_1, \ldots, G_m)$. We model the rare variant contribution to disease risk as $\eta_k = \sum_{j=1}^{m} \beta_{kj} G_j$, $k = 1, \ldots, K$. We assign the variant weights following Wu *et al.* (2011), which are the computed beta distribution density function with parameters 1 and 25 at the rare variant MAF.

We used 10,000 experiments under various combinations of β_{kj} to evaluate the power. We conducted simulations for $\rho = (0.2, 0.5, 0.8)$. For the k-th trait, we set β_{kj} as follows. Each time we randomly selected θ proportion of rare variants and set their $\beta_{kj} = -d \log_{10}(p_j)$, where p_j is the rare variant MAF. The other null rare variants have zero coefficients. We have assumed that rarer variants have larger effect sizes. We conducted simulations for (1) $\theta = 0.25, d = 0.25$, (2) $\theta = 0.5, d = 0.2$, (3) $\theta = 0.75, d = 0.15$. They correspond to regression coefficients of 0.5, 0.4 and 0.3 for MAF=0.01 respectively. We conducted simulations for four scenarios: each time only the first L traits were associated with the rare variant set, $L = 1, 2, 3, 4$. Intuitively in the first scenario ($L = 1$), where only the first trait is associated with the rare variant set, we expect that the minimum p-value based approach or testing the first trait alone will have good performance. But we will show that by simultaneously testing correlated null traits, the proposed MSKAT could actually improve the detection power compared to testing the first trait alone. When there are multiple correlated traits that are associated with the rare variant set, the proposed MSKAT could offer much improved detection power than the minimum p-value based approach.

Table 1, 2, 3, and 4 summarize the power under significance level $\alpha = 10^{-4}$ for $L = 1, 2, 3, 4$ respectively. When only the first trait is associated with the rare variant set (Table 1), Pmin performs better than MKMR, Q and Q' under weak trait correlation ($\rho = 0.2$). Both MKMR and the MSKAT statistic Q could benefit from increased trait correlations, and offer much improved power by incorporating strongly correlated null traits. The statistic Q' ignored the trait dependence by directly summing over individual trait SKAT statistics. Overall we can see that it suffered power loss with increasing trait correlations. The minimum p-value based approach Pmin had nearly constant power across different trait correlations.

When there are multiple correlated traits that are associated with the rare variant set (Table 2, 3, and 4), the MSKAT statistic Q had the overall best performance. Overall we can see that Q' had reduced power with increasing trait correlations, and the Pmin had nearly constant power across different trait correlations. Both MKMR and the MSKAT statistic Q accounted

Table 1: Power of multivariate tests for four continuous traits with pairwise correlation ρ: Q is the proposed MSKAT statistic incorporating the trait correlation, Q' is the sum of individual trait SKAT statistics, Pmin is the Bonferroni corrected minimum p-value based on the individual trait SKAT significance p-values, and MKMR is the multivariate kernel machine regression approach. Only the first trait is associated with the rare variant set ($L = 1$). The causal rare variant proportion is θ and their regression coefficient is set as $-d\log_{10}(\text{MAF})$. The highest powered tests in each column are bold-faced.

(d, θ)	(0.25,0.25)			(0.2,0.5)			(0.15,0.75)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
Q	0.037	**0.070**	**0.257**	0.084	**0.141**	**0.416**	0.088	**0.142**	**0.395**
Q'	0.022	0.008	0.001	0.004	0.026	0.013	0.058	0.029	0.015
Pmin	**0.048**	0.048	0.049	**0.102**	0.102	0.101	**0.104**	0.104	0.105
MKMR	0.010	0.028	0.139	0.028	0.068	0.251	0.032	0.072	0.245

Table 2: Power of multivariate tests for four continuous traits with pairwise correlation ρ: Q is the proposed MSKAT statistic incorporating the trait correlation, Q' is the sum of individual trait SKAT statistics, Pmin is the Bonferroni corrected minimum p-value based on the individual trait SKAT significance p-values, and MKMR is the multivariate kernel machine regression approach. Only the first $L = 2$ traits are associated with the rare variant set. The causal rare variant proportion is θ and their regression coefficient is set as $-d\log_{10}(\text{MAF})$. The highest powered tests in each column are bold-faced.

(d, θ)	(0.25,0.25)			(0.2,0.5)			(0.15,0.75)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
Q	**0.194**	**0.328**	**0.749**	**0.327**	**0.483**	**0.896**	**0.296**	**0.414**	**0.851**
Q'	0.144	0.070	0.012	0.267	0.153	0.019	0.258	0.156	0.102
Pmin	0.180	0.178	0.123	0.302	0.297	0.174	0.277	0.272	0.268
MKMR	0.152	0.293	0.721	0.259	0.440	0.879	0.237	0.380	0.850

for the trait dependence, and had improved power with increasing trait correlations.

When most traits have similar genetic effects (Table 3 and 4), the Q' performed slightly better than Q under weak trait correlation ($\rho = 0.2$).

Overall we can see that the proposed MSKAT statistic Q is an attractive approach with good power across a wide range of alternatives.

Table 5 compared the SKAT based rare variant set testing of Y_1 alone versus the joint multivariate testing under previous simulation settings. We can see that jointly testing highly correlated traits could have greater power over testing Y_1 alone. In general both MKMR and the proposed MSKAT statistic Q could benefit from the trait correlations to largely improve the detection power. The minimum p-value based approach is largely unaffected by the trait

Table 3: Power of multivariate tests for four continuous traits with pairwise correlation ρ: Q is the proposed MSKAT statistic incorporating the trait correlation, Q' is the sum of individual trait SKAT statistics, Pmin is the Bonferroni corrected minimum p-value based on the individual trait SKAT significance p-values, and MKMR is the multivariate kernel machine regression approach. Only the first $L = 3$ traits are associated with the rare variant set. The causal rare variant proportion is θ and their regression coefficient is set as $-d \log_{10}(\text{MAF})$. The highest powered tests in each column are bold-faced.

(d, θ)	(0.25,0.25)			(0.2,0.5)			(0.15,0.75)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
Q	**0.394**	**0.575**	**0.934**	**0.560**	**0.710**	**0.979**	0.464	**0.548**	**0.921**
Q'	0.324	0.173	0.099	0.525	0.326	0.210	**0.483**	0.309	0.209
Pmin	0.275	0.269	0.263	0.417	0.406	0.393	0.367	0.354	0.337
MKMR	0.341	0.552	0.933	0.469	0.664	0.978	0.355	0.473	0.920

Table 4: Power of multivariate tests for four continuous traits with pairwise correlation ρ: Q is the proposed MSKAT statistic incorporating the trait correlation, Q' is the sum of individual trait SKAT statistics, Pmin is the Bonferroni corrected minimum p-value based on the individual trait SKAT significance p-values, and MKMR is the multivariate kernel machine regression approach. All $L = 4$ traits are associated with the rare variant set. The causal rare variant proportion is θ and their regression coefficient is set as $-d \log_{10}(\text{MAF})$. The highest powered tests in each column are bold-faced.

(d, θ)	(0.25,0.25)			(0.2,0.5)			(0.15,0.75)		
ρ	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
Q	**0.581**	**0.739**	**0.975**	0.726	**0.798**	**0.984**	0.573	**0.523**	**0.813**
Q'	0.522	0.312	0.185	**0.739**	0.518	0.351	**0.673**	0.468	0.328
Pmin	0.353	0.343	0.331	0.503	0.485	0.462	0.429	0.406	0.379
MKMR	0.512	0.692	0.972	0.593	0.670	0.972	0.360	0.255	0.681

Table 5: Detection power incorporating correlated null traits: only the first trait Y_1 is associated with the rare variant set. The causal rare variant proportion is θ and their regression coefficient is set as $-d\log_{10}(\text{MAF})$. We compared the multivariate trait based test approach, MKMR, Q, Q' and Pmin, to the SKAT applied to testing Y_1 only, denoted as SKAT(Y_1). The highest powered tests in each row are bold-faced.

$\alpha = 10^{-4}, d = -0.25, \theta = 0.25$					
ρ	SKAT(Y_1)	Q	Q'	Pmin	MKMR
0.2	**0.038**	0.016	0.008	0.024	0.003
0.5	**0.038**	0.038	0.003	0.024	0.011
0.8	0.038	**0.204**	0.001	0.024	0.091
$\alpha = 10^{-4}, d = -0.2, \theta = 0.5$					
ρ	SKAT(Y_1)	Q	Q'	Pmin	MKMR
0.2	**0.054**	0.024	0.012	0.035	0.004
0.5	**0.054**	0.054	0.004	0.035	0.016
0.8	0.054	**0.280**	0.002	0.035	0.128
$\alpha = 10^{-4}, d = -0.15, \theta = 0.75$					
ρ	SKAT(Y_1)	Q	Q'	Pmin	MKMR
0.2	**0.036**	0.017	0.008	0.022	0.002
0.5	**0.036**	0.036	0.002	0.022	0.010
0.8	0.036	**0.205**	0.001	0.023	0.089

correlations.

References

Schaffner,S.F., Foo,C., Gabriel,S., Reich,D., Daly,M.J. and Altshuler,D. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Research,* **15** (11), 1576–1583.

Wu,M., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics,* **89** (1), 82–93.