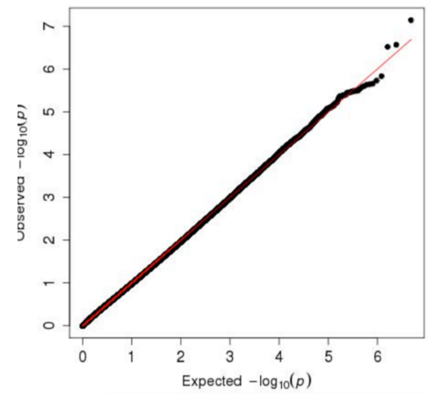
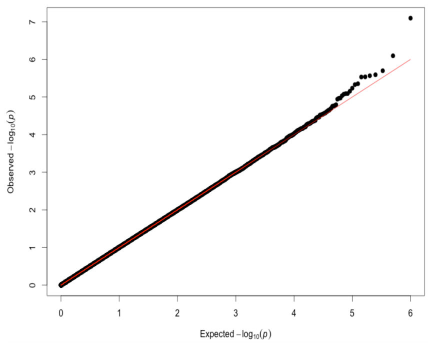
Supplementary Material

Rare variant association test in family based sequencing studies

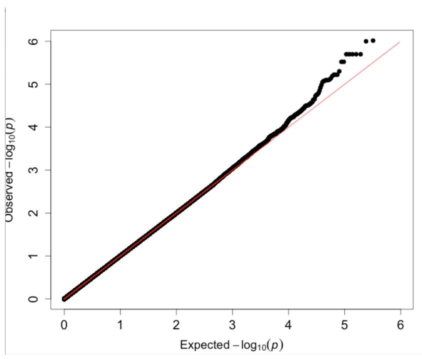
Xuefeng Wang, Zhenyu Zhang, Nathan Morris, Tianxi Cai, Seunggeun Lee, Chaolong Wang, Timothy W. Yu, Christopher A. Walsh, Xihong Lin



(a)

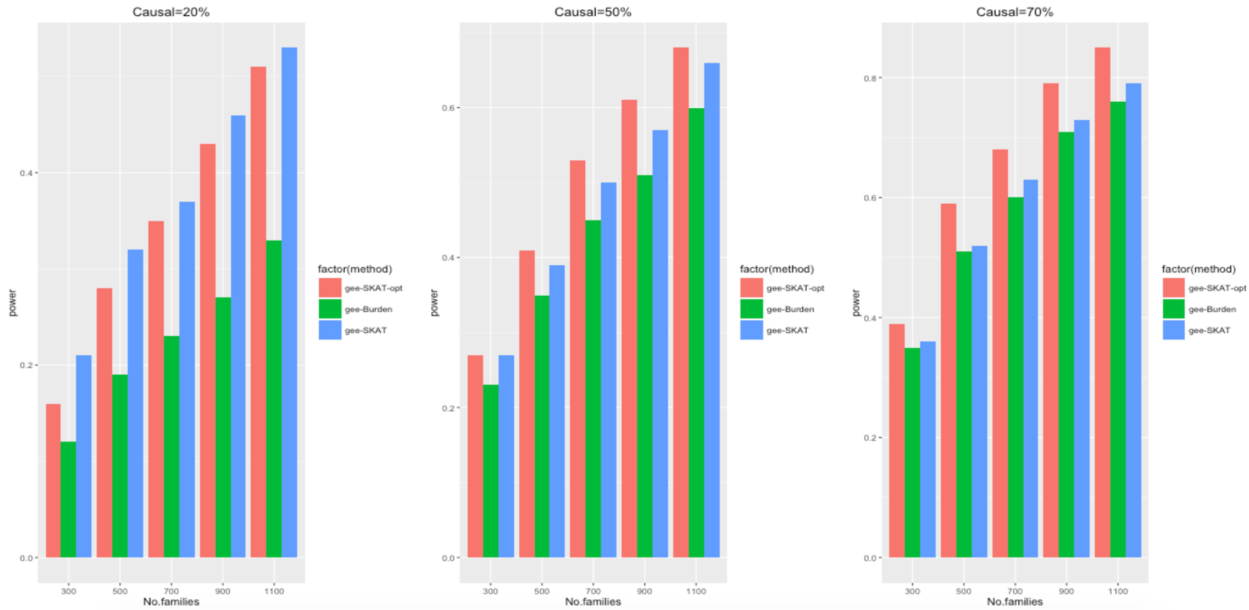


(b)

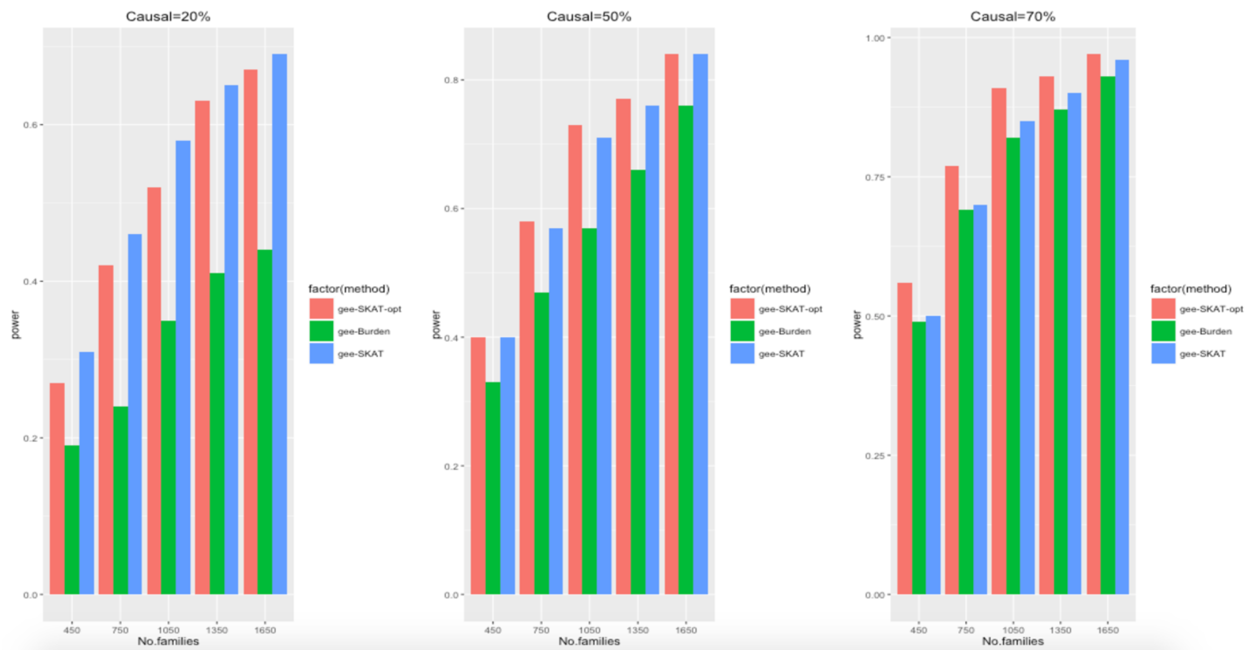


(c)

Supplementary Figure 1: QQ plot through permutation (a) and optimal test (b) from simulations under the null. (c) shows the results from simulation with a mixed sample of sib pairs and nuclear families.



Supplementary Figure 2: Power comparison with sib pair samples, in which half of the families have at least 1 sick child, while other families are unaffected.



Supplementary Figure 3: Power comparison with a mixed sample of sib pairs and nuclear families.

1 Null distribution of the GEE-KM statistic T

To derive the asymptotic distribution of the statistic T, as shown in equation (4), denote $\mathbf{A} = \mathbf{I}(\boldsymbol{\theta_0}) = -E(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\theta}^T}) = \sum_{i=1}^{n} \mathbf{D_i^T V_i^{-1} D_i}$, where $\boldsymbol{\theta_0 = (\alpha_0, 0)'}$ is the true value of $\boldsymbol{\theta}$. Partition $\mathbf{A}$ as $\mathbf{A_{xx}}, \mathbf{A_{xz}}, \mathbf{A_{zx}}, \mathbf{A_{zz}}$ according to the dimensions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. From a Taylor series expansion, we get $\boldsymbol{\tilde{\alpha} - \alpha_0} = \boldsymbol{A_{xx}^{-1} U_x(\theta_0)} + o_p(1)$, where $\boldsymbol{\tilde{\alpha}}$ is the MLE of $\boldsymbol{\alpha}$ under the null.
From a Taylor expansion of $\mathbf{U_z}(\boldsymbol{\tilde{\theta}})$, where $\boldsymbol{\tilde{\theta}} = (\boldsymbol{\tilde{\alpha}}, 0)'$, we have

$$
\begin{aligned}
\mathbf{U_z}(\boldsymbol{\tilde{\theta}_0}) =& [\mathbf{U_z}(\boldsymbol{\theta}_0) - \mathbf{A_z x}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)] + o_p(1) \\
=& [-\mathbf{A_{zx} A_{xx}^{-1} U_x}(\boldsymbol{\theta_0}) + \mathbf{U_z}(\boldsymbol{\theta_0})] + o_p(\sqrt{n})
\end{aligned}
$$

Let $\mathbf{C} = (-\mathbf{A_{zx} A_{xx}^{-1}}, \mathbf{I})$, then $\mathbf{U_z} \approx CU(\boldsymbol{\theta}_0)$. Denote $\mathbf{B} = E(\mathbf{U}(\boldsymbol{\theta_0})\mathbf{U^T}(\boldsymbol{\theta_0})) = \sum_{i=1}^{n} \mathbf{D_i^T V_i^{-1}} Cov(\boldsymbol{y}_i) \mathbf{V_i^{-1} D_i}$. As $n \to \infty$, $\mathbf{B}^{1/2}\mathbf{U}(\boldsymbol{\theta_0}) \to N(0, \mathbf{I})$ in distribution.

$$
\begin{aligned}
T =& \mathbf{U_z^T W R W U_z} \\
=& \{\mathbf{B}^{1/2}\mathbf{U}(\boldsymbol{\theta}_0)\}^T \{\mathbf{B^{1/2} C^T W R W C B^{1/2}}\} \{\mathbf{B}^{1/2}\mathbf{U}(\boldsymbol{\theta}_0)\} \\
\to & \sum_{k=1}^{p} \lambda_k \chi^2_{k,1}
\end{aligned}
$$

where $(\lambda_1, \ldots, \lambda_p)$ are the eigenvalues of $\mathbf{B^{1/2} C^T W R W C B^{1/2}}$ and $\chi^2_{k,1}$ are independent χ^2_1 random variables. $Cov(\boldsymbol{y}_i)$ in $\mathbf{B}$ is estimated by $\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0)_i\}\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0)_i\}^T$

2 Asymptotic approximation of the optimal-test statistic

Following the original SKAT-O (Lee et al. 2012) derivation, denote $\mathbf{Z} = \mathbf{W C B^{1/2}}$, $\bar{z} = \mathbf{Z} \cdot \mathbf{1}_p / p, \mathbf{M} = \bar{z}(\bar{z}^T \bar{z})^{-1} \bar{z}^T$ and $\mathbf{Y} = \mathbf{B^{-1/2}} \mathbf{U}(\theta_0) \sim N(0, \mathbf{I})$, then

$$
T = \mathbf{Y^T Z}[(\mathbf{1} - \rho)\mathbf{I} + \rho \mathbf{1_p 1_p^T}]\mathbf{Z^T Y} = (\mathbf{1} - \rho)\mathbf{Y^T Z Z^T Y} + \rho p^2 \mathbf{Y^T \bar{z} \bar{z}^T Y}
$$

$$
\begin{aligned}
\mathbf{M Z Z^T M} =& \bar{z}(\bar{z}^T \bar{z})^{-1} \bar{z}^T \mathbf{Z Z^T} \bar{z}(\bar{z}^T \bar{z})^{-1} \bar{z}^T \\
=& \frac{1}{(\bar{z}^T \bar{z})^2} \bar{z}\bar{z}^T \mathbf{Z Z^T} \bar{z}\bar{z}^T = \frac{1}{(\bar{z}^T \bar{z})^2} \bar{z}(\mathbf{Z^T} \bar{z})^T (\mathbf{Z^T} \bar{z}) \bar{z}^T \\
=& \frac{\bar{z}\bar{z}^T}{(\bar{z}^T \bar{z})^2} (\mathbf{Z^T} \bar{z})^T (\mathbf{Z^T} \bar{z}) \\
=& \frac{\bar{z}\bar{z}^T}{(\bar{z}^T \bar{z})^2} \sum_{j=1}^{p} (\bar{z}^T \mathbf{Z}_{\cdot j})^2
\end{aligned}
$$

where $\mathbf{Z}_{\cdot j}$ is the jth column of $\mathbf{Z}$.

$$\begin{aligned}
T =&(\mathbf{1}-\rho)\mathbf{Y^T}\mathbf{Z}\mathbf{Z^T}\mathbf{Y}+\rho p^2\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}\\
=&(\mathbf{1}-\rho)\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})\mathbf{Y}\\
&+\mathbf{2}(\mathbf{1}-\rho)\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Y}\\
&+(\mathbf{1}-\rho)\mathbf{Y^T}(\mathbf{M}\mathbf{Z}\mathbf{Z^T}\mathbf{M}+\mathbf{M}\mathbf{Z}\mathbf{Z^T}-\mathbf{Z}\mathbf{Z^T}\mathbf{M})\mathbf{Y}+p^2\rho\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}
\end{aligned}$$

$$\begin{aligned}
&(\mathbf{1}-\rho)\mathbf{Y^T}(\mathbf{M}\mathbf{Z}\mathbf{Z^T}\mathbf{M}+\mathbf{M}\mathbf{Z}\mathbf{Z^T}-\mathbf{Z}\mathbf{Z^T}\mathbf{M})\mathbf{Y}+p^2\rho\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}\\
=&(1-\rho)\frac{\sum_{j=1}^{p}(\bar{z}^T\mathbf{Z}_{\cdot j})^2}{(\bar{z}^T\bar{z})^2}\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}+\rho p^2\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}\\
=&\gamma(\rho)\frac{1}{\bar{z}^T\bar{z}}\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}
\end{aligned}$$

where $\gamma(\rho)=\rho p^2\bar{z}^T\bar{z}+\frac{1-\rho}{\bar{z}^T\bar{z}}\sum_{j=1}^{p}(\bar{z}^T\mathbf{Z}_{\cdot j})^2$.

Denote $\varepsilon=\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})\mathbf{Y}, \zeta=\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Y}, \eta=\mathbf{Y^T}\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}\mathbf{Y}$, as $\mathbf{M}\bar{z}\bar{z}^T=\bar{z}(\bar{z}^T\bar{z})^{-1}\bar{z}^T\bar{z}\bar{z}^T=\bar{z}\bar{z}^T, (\mathbf{I}-\mathbf{M})\bar{\mathbf{z}}\bar{\mathbf{z}}^{\mathbf{T}}=\mathbf{0}$, ε and η are asymptotically independent (Craig's Theorem) under the null.
Since $\mathbf{Y}=\mathbf{B^{-1/2}}\mathbf{U}(\theta_{\mathbf{0}})\sim N(0,\mathbf{I})$, we have

$$\begin{aligned}
E(\zeta)=&E(\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Y})\\
=&tr((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})=tr(\mathbf{M}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T})\\
=&0\\
var(\zeta)=&var(\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Y})\\
=&var\left\{\frac{1}{2}\mathbf{Y^T}\left[(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}+((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}\right]\mathbf{Y}\right\}\\
=&\frac{1}{4}2tr\left\{\left[(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}+((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}\right]^2\right\}\\
=&tr\left[(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\cdot((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}\right]\\
=&tr(\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M}))\\
cov(\varepsilon,\zeta)=&cov(\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})\mathbf{Y},\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\mathbf{Y})\\
=&\frac{1}{2}cov(\mathbf{Y^T}(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})\mathbf{Y},\mathbf{Y^T}[(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}+((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}]\mathbf{Y})\\
=&tr\left\{(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})[(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}+((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}]\right\}\\
=&tr\left\{(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M}\right\}+tr\left\{(\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}(\mathbf{I}-\mathbf{M})((\mathbf{I}-\mathbf{M})\mathbf{Z}\mathbf{Z^T}\mathbf{M})^{\mathbf{T}}\right\}\\
=&0
\end{aligned}$$

Similarly, we can show that $cov(\eta,\zeta)=0$. Let $\kappa=\varepsilon+2\zeta$, $T=(1-\rho)\kappa+\gamma(\rho)\eta$. ε and η are asymptotically independent under the null, ζ is asymptotically uncorrelated with ε and η. Since the Pearson correlation between η and κ is 0, we can approximate T as sum of two independent variables.