Rare variant association tests for ancestry-matched case-control data based on conditional logistic regression

Shanshan Cheng, Jingjing Lyu, Xian Shi, Kai Wang, Zengmiao Wang, Minghua Deng, Baoluo Sun (stasb@nus.edu.sg), Chaolong Wang (chaolong@hust.edu.cn)

1 Supplementary Note

1.1 Derivation of the CLR-SKAT statistic

Assume there are n matched groups. Following the notation in the main text, let $\mathbf{Y} = (\mathbf{Y}_{1\cdot}, \mathbf{Y}_{2\cdot}, ..., \mathbf{Y}_{n\cdot})^T$, $\mathbf{X} = (\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}, ..., \mathbf{X}_{n\cdot})^T$, $\mathbf{G} = (\mathbf{G}_{1\cdot}, \mathbf{G}_{2\cdot}, ..., \mathbf{G}_{n\cdot})^T$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_{1\cdot}, \boldsymbol{\mu}_{2\cdot}, ..., \boldsymbol{\mu}_{n\cdot})^T$, where for the i^{th} matched group, $\boldsymbol{\mu}_{i\cdot} = \left(\mu_{i1}, ..., \mu_{ik(i)}\right)$ with

$$\mu_{im} = \frac{\exp(x_{im}^T\gamma + g_{im}^T\mathbf{W}\beta)}{\sum_{j=1}^{k(i)}\exp(x_{ij}^T\gamma + g_{ij}^T\mathbf{W}\beta)}, \qquad m = 1, 2, ..., k(i).$$

Here $\mathbf{W}$ is a diagonal weight matrix for the SNPs in a gene. It can be shown from the conditional log-likelihood $L(\gamma, \beta) = \log l(\gamma, \beta)$ based on (3) in the main text that

$$\frac{\partial L}{\partial \gamma} = \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) \tag{A1}$$

$$\frac{\partial L}{\partial \beta} = \mathbf{W}\mathbf{G}^T(\mathbf{Y} - \boldsymbol{\mu}) \tag{A2}$$

$$\frac{\partial^2 L}{\partial \gamma \partial \gamma^T} = -\mathbf{X}^T\mathbf{V}\mathbf{X} \tag{A3}$$

$$\frac{\partial^2 L}{\partial \gamma \partial \beta^T} = -\mathbf{X}^T\mathbf{V}\mathbf{G}\mathbf{W} \tag{A4}$$

$$\frac{\partial^2 L}{\partial \beta \partial \beta^T} = -\mathbf{W}\mathbf{G}^T\mathbf{V}\mathbf{G}\mathbf{W}, \tag{A5}$$

where $\mathbf{V} = \mathrm{diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T$. We assume that β is a $q \times 1$ vector of random effects with mean $\mathbf{1}_q\beta_0$ and covariance matrix $\tau\left\{(1-\rho)\mathbf{I}_q + \rho\mathbf{1}_q\mathbf{1}_q^T\right\}$, $\mathbf{I}_q$ is a $q \times q$ identity matrix and $\mathbf{1}_q$ is a column vector of length q with all elements 1. The conditional log-likelihood with respect to the parameters

$(\gamma, \tau, \beta_0, \rho)$ is

$$\begin{aligned}
\tilde{L}(\gamma, \tau, \beta_0, \rho) &= \log \int \exp\left[L(\gamma, \beta)\right] dF(\beta; \beta_0, \tau, \rho) \\
&= \log \int \exp\left[L(\gamma, \beta_0)\right] \Bigg\{ 1 + \frac{\partial L(\gamma, \beta_0)}{\partial \beta^T} (\beta - \beta_0) \\
&\quad + \frac{1}{2} (\beta - \beta_0)^T \left[\frac{\partial L(\gamma, \beta_0)}{\partial \beta} \frac{\partial L(\gamma, \beta_0)}{\partial \beta^T} + \frac{\partial L^2(\gamma, \beta_0)}{\partial \beta \partial \beta^T} \right] (\beta - \beta_0) \Bigg\} dF(\beta; \beta_0, \tau, \rho) \\
&= L(\gamma, \beta_0) + \log \Bigg\{ 1 + \frac{1}{2} \tau \mathrm{tr} \left[\mathbf{L}^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{L} \right. \\
&\quad \left. - \mathbf{L}^T \mathbf{W} \mathbf{G}^T \mathbf{V} \mathbf{G} \mathbf{W} \mathbf{L} \right] \Bigg\},
\end{aligned}$$

where the second equality above follows from a two-term Taylor expansion around $\beta = \beta_0$ and $\mathbf{L}\mathbf{L}^T = (1-\rho)\mathbf{I}_q + \rho \mathbf{1}_q \mathbf{1}_q^T$ is the Cholesky decomposition of the correlation matrix of the genotype effects β. The score of the variance component τ is

$$\frac{\partial \tilde{L}(\gamma, \tau, \beta_0, \rho)}{\partial \tau} = \frac{\frac{1}{2} \mathrm{tr} \left[\mathbf{L}^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{L} - \mathbf{L}^T \mathbf{W} \mathbf{G}^T \mathbf{V} \mathbf{G} \mathbf{W} \mathbf{L} \right]}{1 + \frac{1}{2} \tau \mathrm{tr} \left[\mathbf{L}^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{L} - \mathbf{L}^T \mathbf{W} \mathbf{G}^T \mathbf{V} \mathbf{G} \mathbf{W} \mathbf{L} \right]}.$$

The CLR-SKAT statistic under $H_0 : \tau = 0$ assuming $\beta_0 = 0$ and $\rho = 0$ is

$$\begin{aligned}
Q &= \mathrm{tr} \left[\mathbf{W} \mathbf{G}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}})(\mathbf{Y} - \hat{\boldsymbol{\mu}})^T \mathbf{G} \mathbf{W} - \mathbf{W} \mathbf{G}^T \hat{\mathbf{V}} \mathbf{G} \mathbf{W} \right] \\
&= (\mathbf{Y} - \hat{\boldsymbol{\mu}})^T \mathbf{G} \mathbf{W} \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}}) - \mathrm{tr} \left(\mathbf{W} \mathbf{G}^T \hat{\mathbf{V}} \mathbf{G} \mathbf{W} \right) \qquad \text{(A6)}
\end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\gamma})$ and $\hat{\mathbf{V}} = \mathbf{V}(\hat{\gamma})$, $\hat{\gamma}$ is the maximum-likelihood estimator under H_0, i.e. $\hat{\gamma} = \arg\max_\gamma L(\gamma, 0)$. Conditional on the unordered set of exposures in each matched group, the second term in (A6) is a constant, and therefore we take the first term as the CLR-SKAT statistic. The score of the mean β_0 is

$$\frac{\partial \tilde{L}(\gamma, \tau, \beta_0, \rho)}{\partial \beta_0} = \frac{\partial L(\gamma, \mathbf{1}_q \beta_0)}{\partial \beta_0} + \frac{\frac{1}{2} \tau \frac{\partial}{\partial \beta_0} \mathrm{tr} \left[\mathbf{L}^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{L} - \mathbf{L}^T \mathbf{W} \mathbf{G}^T \mathbf{V} \mathbf{G} \mathbf{W} \mathbf{L} \right]}{1 + \frac{1}{2} \tau \mathrm{tr} \left[\mathbf{L}^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{L} - \mathbf{L}^T \mathbf{W} \mathbf{G}^T \mathbf{V} \mathbf{G} \mathbf{W} \mathbf{L} \right]}.$$

The CLR-burden statistic under $H_0 : \beta_0 = 0$ assuming $\tau = 0$ is

$$Q_{burden} = \left[\frac{\partial L(\gamma, \mathbf{1}_q \beta_0)}{\partial \beta_0} \right]^2 = (\mathbf{Y} - \hat{\boldsymbol{\mu}})^T \mathbf{G} \mathbf{W} \mathbf{1}_q \mathbf{1}_q^T \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}}),$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\gamma})$ and $\hat{\gamma} = \arg\max_\gamma L(\gamma, 0)$. This is also the variance component test statistic when $\rho = 1$.

1.2 Asymptotic distributions of CLR-SKAT and CLR-burden under H_0

The score $\frac{\partial L}{\partial \beta} = \mathbf{W} \mathbf{G}^T (\mathbf{Y} - \boldsymbol{\mu})$ has asymptotically normal distribution with mean zero and variance given by the information matrix

$$\boldsymbol{\Sigma} = - \left\{ \frac{\partial^2 L}{\partial \beta \partial \beta^T} - \frac{\partial^2 L^T}{\partial \gamma \partial \beta^T} \frac{\partial^2 L^{-1}}{\partial \gamma \partial \gamma^T} \frac{\partial^2 L}{\partial \gamma \partial \beta^T} \right\} = \mathbf{W} \mathbf{G}^T \left\{ \mathbf{V} - \mathbf{V}^T \mathbf{X} (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V} \right\} \mathbf{G} \mathbf{W}.$$

Under H_0 and standard regularity conditions, $\hat{\gamma} \xrightarrow{p} \gamma_0$. The asymptotic distribution of the CLR-SKAT statistic is then

$$\begin{aligned} Q &= (\mathbf{Y}-\hat{\boldsymbol{\mu}})^T \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}}) \\ &= \left[\boldsymbol{\Sigma}^{-1/2}\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right]^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma}^{-1/2}\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right] \\ &= \tilde{y}^T H^T \Psi H \tilde{y} \\ &\sim \sum_{k=1}^{q} \lambda_k \chi^2_{1,k}, \end{aligned} \tag{A7}$$

where $\tilde{y} \sim N(0, I)$, $\lambda_1 \geq ... \geq \lambda_q$ are the ordered non-zero eigenvalues of $\boldsymbol{\Sigma}$, $\Psi = \text{diag}(\lambda_k)$ and H is the matrix of corresponding eigenvectors of λ_k such that $HH^T = I$. $\chi^2_{1,k}, k = 1, 2, ..., q$ follow independent chi-square distributions with 1 degree of freedom. For CLR-burden,

$$\begin{aligned} Q_{burden} &= (\mathbf{Y}-\hat{\boldsymbol{\mu}})^T \mathbf{G}\mathbf{W}\mathbf{1}_q\mathbf{1}_q^T\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}}) \\ &= \left[\left(\mathbf{1}_q^T\boldsymbol{\Sigma}\mathbf{1}_q\right)^{-1/2}\mathbf{1}_q^T\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right]^T \mathbf{1}_q^T\boldsymbol{\Sigma}\mathbf{1}_q \left[\left(\mathbf{1}_q^T\boldsymbol{\Sigma}\mathbf{1}_q\right)^{-1/2}\mathbf{1}_q^T\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right] \\ &\sim \left(\mathbf{1}_q^T\boldsymbol{\Sigma}\mathbf{1}_q\right)\chi^2_1. \end{aligned}$$

1.3 Equivalent representations of CLR-SKAT and Burden

Let $\mathbf{g}_j$ denote the j^{th} column vector of $\mathbf{G}$, so that $\mathbf{g}_j$ represents the genotype information of the j^{th} variant in the gene for all individuals. The CLR-SKAT statistic can be expressed as

$$\begin{aligned} Q &= (\mathbf{Y}-\hat{\boldsymbol{\mu}})^T \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}}) \\ &= \sum_{j=1}^{q} w_j^2 \left[\mathbf{g}_j^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right]^2 \\ &= \sum_{j=1}^{q} \left\{ w_j^2 \mathbf{g}_j^T \left[\hat{\mathbf{V}} - \hat{\mathbf{V}}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}\right]\mathbf{g}_j \right\} \times \left\{ \frac{\mathbf{g}_j^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})}{\sqrt{\mathbf{g}_j^T \left[\hat{\mathbf{V}} - \hat{\mathbf{V}}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}\right]\mathbf{g}_j}} \right\}^2 \\ &= \sum_{j=1}^{q} \sigma_{jj} \times z_j^2, \end{aligned}$$

where σ_{jj} is the j^{th} diagonal element of $\boldsymbol{\Sigma}$ and

$$z_j = \frac{\mathbf{g}_j^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})}{\sqrt{\mathbf{g}_j^T \left\{\hat{\mathbf{V}} - \hat{\mathbf{V}}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}\right\}\mathbf{g}_j}},$$

which is the standardized score statistic of the marginal test for the significance of the j^{th} variant in the gene under conditional logistic model. The score statistic for the burden test is given by

$$\begin{aligned}
Q_{burden} &= (\mathbf{Y}-\hat{\boldsymbol{\mu}})^T\mathbf{G}\mathbf{W}\mathbf{1}_q\mathbf{1}_q^T\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}}) \\
&= \left[\sum_{j=1}^{q} w_j\mathbf{g}_j^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})\right]^2 \\
&= \left\{\sum_{j=1}^{q}\left[w_j\sqrt{\mathbf{g}_j^T\left[\hat{\mathbf{V}}-\hat{\mathbf{V}}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}\right]\mathbf{g}_j}\right]\times\left[\frac{\mathbf{g}_j^T(\mathbf{Y}-\hat{\boldsymbol{\mu}})}{\sqrt{\mathbf{g}_j^T\left[\hat{\mathbf{V}}-\hat{\mathbf{V}}^T\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}\right]\mathbf{g}_j}}\right]\right\}^2 \\
&= \left\{\sum_{j=1}^{q}\sqrt{\sigma_{jj}}\times z_j\right\}^2.
\end{aligned}$$

1.4 CLR-MiST

For CLR-MiST, besides CLR-burden we construct another asymptotically independent test under $H_0:\tau=0$ with $\rho=0$ and β_0 unrestricted. This SKAT-type statistic is

$$Q_{\tilde{\beta}} = (\mathbf{Y}-\hat{\boldsymbol{\mu}}_{\tilde{\beta}})^T\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T(\mathbf{Y}-\hat{\boldsymbol{\mu}}_{\tilde{\beta}}),$$

where $\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}\left(\hat{\gamma},\tilde{\beta}\right)$ and $\left(\hat{\gamma},\tilde{\beta}\right)$ is the maximum-likelihood estimator under H_0, i.e.

$$\left(\hat{\gamma},\tilde{\beta}\right) = \arg\max_{\gamma,\beta_0} L(\gamma,\mathbf{1}_q\beta_0).$$

Following section 1.3, $Q_{\tilde{\beta}}$ has an equivalent representation as

$$Q_{\tilde{\beta}} = \sum_{j=1}^{q}\tilde{\sigma}_{jj}\times\tilde{z}_j^2,$$

where $\tilde{z}$ is the standardized score statistic of the marginal test for the significance of the j^{th} variant in the gene under conditional logistic null mode with the genetic burden included as a fixed effects covariate, and $\tilde{\sigma}_{jj}$ is the j^{th} diagonal element of $\tilde{\boldsymbol{\Sigma}}=\mathbf{W}\mathbf{G}^T\left\{\tilde{\mathbf{V}}-\tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\right\}\mathbf{G}\mathbf{W}$, where $\tilde{\mathbf{X}}=(\mathbf{X},\mathbf{G}\mathbf{W}\mathbf{1}_q)$. Following section 1.2,

$$Q_{\tilde{\beta}} \sim \sum_{k=1}^{q}\tilde{\lambda}_k\chi_{1,k}^2,$$

where $\tilde{\lambda}_1\geq...\geq\tilde{\lambda}_q$ are the ordered non-zero eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ and $\chi_{1,k}^2, k=1,2,...,q$ follow independent chi-square distributions with 1 degree of freedom. We adapt the proof from the Appendix of Sun et al. (2013) to show that Q_{burden} and $Q_{\tilde{\beta}}$ are asymptotically independent under $H_0:\tau=0,\beta_0=0$. At convergence of Fisher scoring,

$$\mathbf{Y} - \hat{\boldsymbol{\mu}} \approx \left\{\mathbf{I} - \mathbf{V}^T\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\right\}(\mathbf{Y} - \boldsymbol{\mu}_0) = P_1(\mathbf{Y} - \boldsymbol{\mu}_0)$$
$$\mathbf{Y} - \hat{\boldsymbol{\mu}}_{\tilde{\beta}} \approx \left\{\mathbf{I} - \tilde{\mathbf{V}}^T\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\right\}(\mathbf{Y} - \boldsymbol{\mu}_0) = P_2(\mathbf{Y} - \boldsymbol{\mu}_0).$$

Let $U_{burden} = \mathbf{1}_q^T\mathbf{W}\mathbf{G}^T P_1(\mathbf{Y} - \boldsymbol{\mu}_0)$, $U_{\tilde{\beta}} = \mathbf{W}\mathbf{G}^T P_2(\mathbf{Y} - \boldsymbol{\mu}_0)$ and $V^T = \left(U_{burden}^T, U_{\tilde{\beta}}^T\right)$. By central limit theorem, V converges to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_V. The off-diagonal matrix of Σ_V is

$$\begin{aligned}
\mathrm{Cov}\left(U_{burden}, U_{\tilde{\beta}}\right) &= \mathrm{Cov}\left\{\mathbf{1}_q^T\mathbf{W}\mathbf{G}^T P_1(\mathbf{Y} - \boldsymbol{\mu}_0), \mathbf{W}\mathbf{G}^T P_2(\mathbf{Y} - \boldsymbol{\mu}_0)\right\} \\
&= \mathbf{1}_q^T\mathbf{W}\mathbf{G}^T P_1 \mathrm{Cov}(\mathbf{Y} - \boldsymbol{\mu}_0, \mathbf{Y} - \boldsymbol{\mu}_0) P_2^T\mathbf{G}\mathbf{W}^T \\
&= \mathbf{1}_q^T\mathbf{W}\mathbf{G}^T P_1 V P_2^T\mathbf{G}\mathbf{W}^T \\
&= \mathbf{1}_q^T\mathbf{W}\mathbf{G}^T P_2 V\mathbf{G}\mathbf{W}^T - \mathbf{1}_q^T\mathbf{W}\mathbf{G}^T\mathbf{V}^T\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T P_2 V \\
&= 0,
\end{aligned}$$

where $\tilde{\mathbf{X}}^T P_2 V = 0$ and $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{G}\mathbf{W}\mathbf{1}_q)$. Since $\left(U_{burden}^T, U_{\tilde{\beta}}^T\right)$ are asymptotically jointly normal, Q_{burden} and $Q_{\tilde{\beta}}$ are asymptotically independent. Let p_{burden} and $p_{\tilde{\beta}}$ be the p-values of the two tests respectively. By Fisher's procedure, CLR-MiST rejects H_0 at an overall significance level of α if

$$-2\log(p_{burden}) - 2\log(p_{\tilde{\beta}}) \geq \chi^2_{4,\alpha}.$$

1.5 Extension to arbitrary numbers of cases and controls per matched group

Consider the situation with $r(i)$ cases and $s(i)$ controls the i^{th} in the matched group, with $k(i) = r(i) + s(i)$ subjects in total, $i = 1, 2, ..., n$. In order to condition on the unordered set of $k(i)$ exposures, there are $\binom{k(i)}{r(i)}$ ways of partitioning the matched group into $r(i)$ cases and $s(i)$ controls. For each such partition $\mathcal{D}$, let the indices assigned to cases be $\mathcal{D}(1), ..., \mathcal{D}(r(i))$. Analogous to (3) in the main text, the conditional likelihood is given by

$$l(\gamma, \beta) = \prod_{i=1}^{n} \frac{\exp\left\{\sum_{j=1}^{k(i)}\left[y_{ij}\left(x_{ij}^T\gamma + g_{ij}^T\mathbf{W}\beta\right)\right]\right\}}{\sum_{\mathcal{D}}\exp\left\{\sum_{j=\mathcal{D}(1)}^{\mathcal{D}(r(i))}\left(x_{ij}^T\gamma + g_{ij}^T\mathbf{W}\beta\right)\right\}}, \tag{A8}$$

where $\sum_{\mathcal{D}}$ refers to summation over all the $\binom{k(i)}{r(i)}$ partitions in the i^{th} matched group. For the i^{th} matched group, let $\mathbf{K}_i$ be the $\binom{k(i)}{r(i)} \times k(i)$ matrix of assigned case indicators, with each row vector consisting of 1's at the indices $\mathcal{D}(1), ..., \mathcal{D}(r(i))$ and 0's otherwise. Let $\mathbf{L}_i$ be the corresponding $1 \times \binom{k(i)}{r(i)}$ vector of conditional probabilities of observing the exposures assigned as cases in each partition of the i^{th} matched group, i.e. each entry is

$$\frac{\exp\left\{\sum_{j=\tilde{\mathcal{D}}(1)}^{\tilde{\mathcal{D}}(r(i))}\left(x_{ij}^T\gamma + g_{ij}^T\mathbf{W}\beta\right)\right\}}{\sum_{\mathcal{D}}\exp\left\{\sum_{j=\mathcal{D}(1)}^{\mathcal{D}(r(i))}\left(x_{ij}^T\gamma + g_{ij}^T\mathbf{W}\beta\right)\right\}},$$

for a specific partition $\tilde{\mathcal{D}}$. Let $\tilde{\boldsymbol{\mu}} = (\mathbf{L}_1\mathbf{K}_1, \mathbf{L}_2\mathbf{K}_2, ..., \mathbf{L}_n\mathbf{K}_n)^T$. Then analogous to (A1)-(A5), we have

$$\frac{\partial L}{\partial \gamma} = \mathbf{X}^T(\mathbf{Y} - \tilde{\boldsymbol{\mu}}) \tag{A9}$$

$$\frac{\partial L}{\partial \beta} = \mathbf{W}\mathbf{G}^T(\mathbf{Y} - \tilde{\boldsymbol{\mu}}) \tag{A10}$$

$$\frac{\partial^2 L}{\partial \gamma \partial \gamma^T} = -\mathbf{X}^T\tilde{\mathbf{V}}\mathbf{X} \tag{A11}$$

$$\frac{\partial^2 L}{\partial \gamma \partial \beta^T} = -\mathbf{X}^T\tilde{\mathbf{V}}\mathbf{G}\mathbf{W} \tag{A12}$$

$$\frac{\partial^2 L}{\partial \beta \partial \beta^T} = -\mathbf{W}\mathbf{G}^T\tilde{\mathbf{V}}\mathbf{G}\mathbf{W}, \tag{A13}$$

where $\tilde{\mathbf{V}} = \text{diag}(\tilde{\boldsymbol{\mu}}) - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T$. The SKAT statistic is similarly derived as

$$Q = [\mathbf{Y} - \tilde{\boldsymbol{\mu}}(\hat{\gamma})]^T\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T[\mathbf{Y} - \tilde{\boldsymbol{\mu}}(\hat{\gamma})] \tag{A14}$$

where $\hat{\gamma}$ is obtained by maximizing (A8) under H_0. In practice, the Efron or Breslow's approximation may be used when the number of partitions is large and fitting the exact conditional likelihood (A8) tends to be computationally demanding.

1.6 Models and parameter settings for type 1 error and power simulations

Type 1 error simulation was carried out under two case sampling schemes. Let U denote the grid coordinates of the 20×20 lattice which takes value in the set $\{(c_1, c_2) : c_1 \in \{0, 1, 2, ..., 19\}, c_2 \in \{0, 1, 2, ..., 19\}\}$. For sharp spatial risk distribution, cases are sampled with probability

$$\Pr(Y = 1|U = (u_1, u_2), G = g) = \frac{\exp[\theta_0 + \theta_1 \mathbb{1}((u_1, u_2) \in \mathbb{L})]}{1 + \exp[\theta_0 + \theta_1 \mathbb{1}((u_1, u_2) \in \mathbb{L})]},$$

where $\mathbb{1}(\cdot)$ is the indicator variable, $\mathbb{L} = \{(c_1, c_2) : c_1 \in \{10, 11, ..., 19\}, c_2 \in \{10, 11, ..., 19\}\}$ denotes the set of grid coordinates in the top right 7×7 grids of the 20×20 lattice, and $\theta = (\theta_0, \theta_1) = (-3, 2.6)$ in the simulation. The parameters θ are chosen so that disease prevalence is $\approx 40\%$ in the region of high risk and $\approx 5\%$ in the region of low risk to ensure enough cases. For smooth spatial risk distribution, cases are sampled with probability

$$\Pr(Y = 1|U = (u_1, u_2), G = g) = \frac{\exp\left\{\tilde{\theta}_0 + \tilde{\theta}_1\left[(u_1 - 14)^2 + (u_2 - 14)^2\right]\right\}}{1 + \exp\left\{\tilde{\theta}_0 + \tilde{\theta}_1\left[(u_1 - 14)^2 + (u_2 - 14)^2\right]\right\}},$$

with $\tilde{\theta} = \left(\tilde{\theta}_0, \tilde{\theta}_1\right) = (-0.5, -0.03)$ in the simulation. Therefore the disease prevalence decays from a maximum of $\approx 38\%$ at the upper right grid coordinates $(14, 14)$ to a minimum of $\approx 0\%$ at the lower left corner grid coordinates $(0, 0)$. We estimated type 1 error at $\alpha = 10^{-3}$ and 2.5×10^{-6} based on 10^7 association tests.

For power simulation under sharp spatial risk distribution, cases are sampled with probability

$$\Pr(Y = 1|U = (u_1, u_2), G = g) = \frac{\exp[\theta_0 + \theta_1 \mathbb{1}((u_1, u_2) \in \mathbb{L}) + g^T\beta]}{1 + \exp[\theta_0 + \theta_1 \mathbb{1}((u_1, u_2) \in \mathbb{L}) + g^T\beta]},$$

where $g = (g_1, g_2, ..., g_q)^T$ denotes the vector of q genotypes with effects $\beta = (\beta_1, \beta_2, ..., \beta_q)^T$ on the disease risk and $(\tilde{\theta}_0, \tilde{\theta}_1) = (-3, 2.6)$ in the simulation. We randomly selected 20% or 50% of the variants in each gene to be causal, setting the effects for the remaining non-causal variants to be zero. Of the causal variants, the effect sizes are set as $|\beta_j| = c|\log_{10} MAF_j|/2$ for the j^{th} variant, so that rarer variants exert bigger influences on disease risk. To distinguish the powers, for causal variants we set $c = 0.877, 0.555$ when 20% and 50% of the variants are causal respectively. To investigate settings with different effect directions, 80% or 100% of the causal variants are randomly selected to be deleterious (negative β_j), and the rest protective (positive β_j). Under the smooth spatial risk distribution, cases are sampled with probability

$$\Pr(Y = 1|U = (u_1, u_2), G = g) = \frac{\exp\left\{\tilde{\theta}_0 + \tilde{\theta}_1\left[(u_1 - 14)^2 + (u_2 - 14)^2\right] + g^T\beta\right\}}{1 + \exp\left\{\tilde{\theta}_0 + \tilde{\theta}_1\left[(u_1 - 14)^2 + (u_2 - 14)^2\right] + g^T\beta\right\}},$$

where $(\tilde{\theta}_0, \tilde{\theta}_1) = (-0.5, -0.03)$ in the simulation. Empirical power is evaluated as the number of association tests attaining significance at 2.5×10^{-6} out of 1000 replicates.

1.7 Caliper width estimation

The variances of PC1 and PC2 are estimated within each of the 400 grids. Then the estimated are combined via a weighted approach, in which the weights are inversely proportional to the number of samples in each grid, i.e. the standard error of the p^{th} PC is given by

$$\hat{\sigma}_p = \sqrt{\sum_{j=1}^{400}\left\{\frac{n_i}{\sum_{k=1}^{400} n_k}\hat{\sigma}^2_{j,p}\right\}},$$

where n_i is the number of samples in the i^{th} grid, which is a constant of 50 for all grids in the simulation. Using this approach, the approximate 95% prediction width is given by

$$\hat{w}_{0.95} = \sqrt{\chi^2(1 - 0.05, 2)\frac{\hat{\sigma}_1 + \hat{\sigma}_2}{2}} \approx 6.05,$$

which we set as the caliper width in the simulation study.

References

Jianping Sun, Yingye Zheng, and Li Hsu. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, 37(4):334–344, 2013.

Supplementary Tables and Figures

Table S1. Empirical relative type 1 errors for simulations without control down-sampling.

Spatial risk	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
Smooth	1-to-1	0.24 (0.37)	1.48 (1.05)	0.40 (0.60)	1.04 (1.03)	0.32 (0.39)	1.24 (1.05)
	1-to-3	0.52 (0.43)	1.24 (1.09)	0.36 (0.62)	1.00 (1.03)	0.24 (0.41)	1.40 (1.07)
	Full	38.9 (7.50)	1.16 (1.10)	10.2 (2.99)	1.16 (1.03)	27.5 (7.07)	1.24 (1.09)
Sharp	1-to-1	0.24 (0.32)	1.28 (1.03)	0.12 (0.57)	0.56 (1.03)	0.24 (0.36)	1.00 (1.06)
	1-to-3	0.24 (0.39)	1.36 (1.09)	0.16 (0.57)	1.00 (1.01)	0.28 (0.36)	1.48 (1.07)
	Full	2717 (100)	1.36 (1.08)	191 (14.8)	0.60 (1.00)	2462 (101)	1.08 (1.07)

Each entry represents the proportion of significant tests out of 10 million simulation replicates at $\alpha=2.5\times10^{-6}$ or 10^{-3} (in the parentheses), divided by α.

Table S2. Empirical power for the smooth risk setting without control down-sampling.

Effect direction	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
4/16/80 (+/-/0, %)	1-to-1	0.268	0.341	0.116	0.133	0.347	0.421
	1-to-3	0.574	0.589	0.248	0.257	0.604	0.662
	Full	-	0.652	-	0.303	-	0.725
0/20/80 (+/-/0, %)	1-to-1	0.314	0.387	0.208	0.243	0.448	0.523
	1-to-3	0.627	0.643	0.403	0.419	0.725	0.759
	Full	-	0.703	-	0.487	-	0.804
10/40/50 (+/-/0, %)	1-to-1	0.299	0.367	0.225	0.262	0.418	0.485
	1-to-3	0.565	0.580	0.426	0.439	0.669	0.718
	Full	-	0.655	-	0.502	-	0.781
0/50/50 (+/-/0, %)	1-to-1	0.385	0.451	0.518	0.564	0.602	0.665
	1-to-3	0.666	0.672	0.748	0.759	0.819	0.849
	Full	-	0.749	-	0.807	-	0.902

Each entry represents the proportion of significant tests out of 1,000 replicates at $\alpha=2.5\times10^{-6}$. Dashes indicate settings with inflated type 1 error.

Table S3. Empirical power for the sharp risk setting without control down-sampling.

Effect direction	Matching scheme	SKAT	CLR-SKAT	Burden	CLR-Burden	MiST	CLR-MiST
4/16/80 (+/-/0, %)	1-to-1	0.290	0.381	0.135	0.154	0.366	0.455
	1-to-3	0.545	0.577	0.249	0.259	0.596	0.657
	Full	-	0.674	-	0.319	-	0.757
0/20/80 (+/-/0, %)	1-to-1	0.357	0.450	0.241	0.283	0.490	0.595
	1-to-3	0.641	0.659	0.425	0.451	0.744	0.804
	Full	-	0.753	-	0.532	-	0.884
10/40/50 (+/-/0, %)	1-to-1	0.318	0.398	0.251	0.283	0.432	0.519
	1-to-3	0.534	0.562	0.411	0.435	0.647	0.723
	Full	-	0.655	-	0.517	-	0.799
0/50/50 (+/-/0, %)	1-to-1	0.411	0.503	0.557	0.628	0.638	0.734
	1-to-3	0.658	0.676	0.775	0.801	0.847	0.889
	Full	-	0.771	-	0.864	-	0.937

Each entry represents the proportion of significant tests out of 1,000 replicates at $\alpha=2.5\times10^{-6}$. Dashes indicate settings with inflated type 1 error.

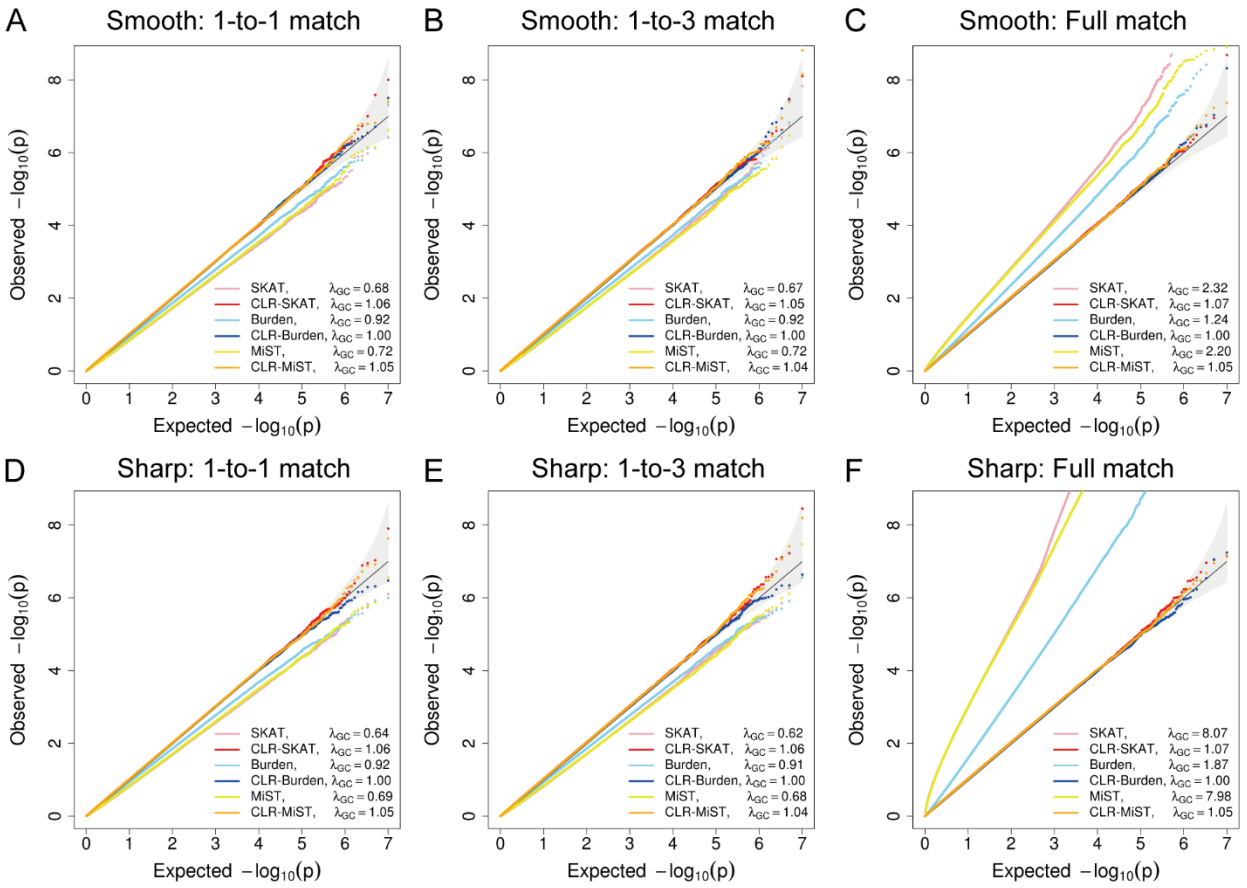

Figure S1. QQ plots for evaluating type 1 errors under the null model without control down-sampling. (A) 1-to-1 match under the smooth risk setting. (B) 1-to-3 match under the smooth risk setting. (C) Full match under the smooth risk setting. (D) 1-to-1 match under the sharp risk setting. (E) 1-to-3 match under the sharp risk setting. (F) Full match under the sharp risk setting. The genomic inflation factor λ_{GC} for each test was indicated in the legend.

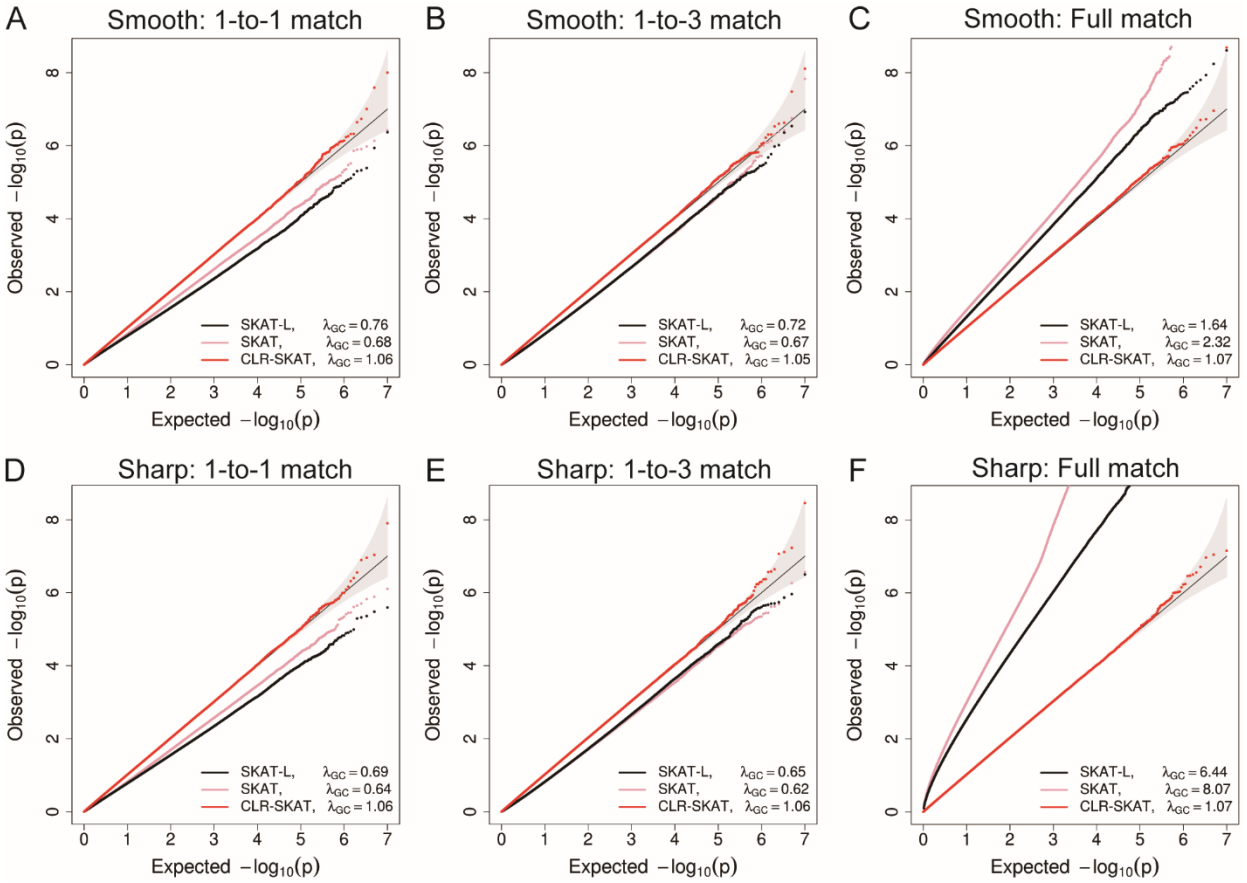

Figure S2. QQ plots comparing SKAT-L, SKAT and CLR-SKAT under the null model without control down-sampling. (A) 1-to-1 match under the smooth risk setting. (B) 1-to-3 match under the smooth risk setting. (C) Full match under the smooth risk setting. (D) 1-to-1 match under the sharp risk setting. (E) 1-to-3 match under the sharp risk setting. (F) Full match under the sharp risk setting. The genomic inflation factor λ_{GC} for each test was indicated in the legend.

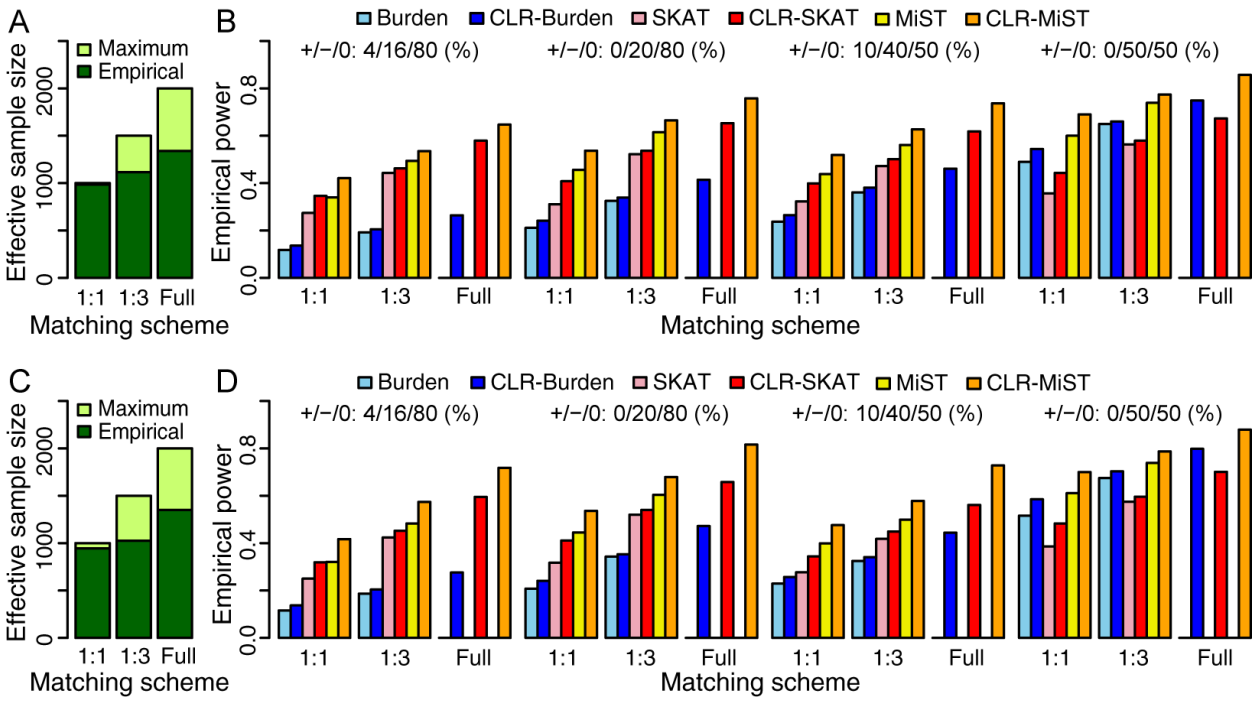

Figure S3. Statistical power evaluated at $\alpha=2.5\times10^{-6}$ based on 1,000 simulation replicates without control down-sampling. (A) Effective sample size under smooth risk setting. Empirical values are calculated as the average effective sample size across 1,000 simulation replicates; maximum values indicate the theoretical upper limit. (B) Empirical power under smooth risk for different settings of effect directions for the causal variants (C) Effective sample size under sharp risk setting. (D) Empirical power under sharp risk for different settings of effect directions for the causal variants. In (B) and (D), +/-/0 represent the percentages of protective, deleterious and non-causal variants, respectively.

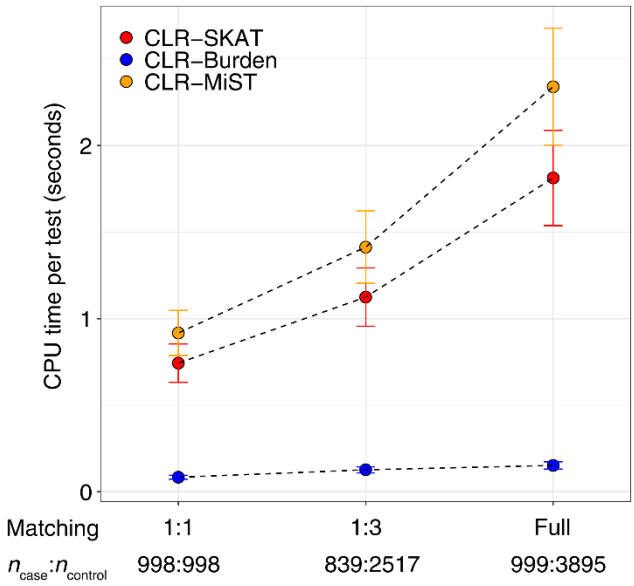

Figure S4. Computational time of CLR-based tests. The evaluation was based on the smooth risk setting null simulations with control down-sampling (corresponding to **Figures 3A-C**). We computed the mean computational time per test, averaging across 10,000 genes. Error bar indicates one standard error of the mean. The actual sample sizes of cases (n_{case}) and controls ($n_{control}$) for 1-to-1, 1-to-3, and full matching in this experiment were presented in the bottom. Computation was performed on a single thread of a 2.0 GHz CPU processor (Intel Xeon Gold 6138).