Original Paper

Human Heredity

Hum Hered 2016;82:50–63
DOI: 10.1159/000478048

Received: January 30, 2017
Accepted: June 7, 2017
Published online: August 16, 2017

Efficient Statistical Method for Association Analysis of X-Linked Variants

Heejin Jin[a] Taesung Park[b, c] Sungho Won[a, c, d]

[a]Department of Public Health Science, Graduate School of Public Health, [b]Department of Statistics, and [c]Interdisciplinary Program in Bioinformatics, College of Natural Science, and [d]Institute of Health and Environment, Seoul National University, Seoul, Korea

Keywords
X-chromosome inactivation · X-linked variants · X-chromosome association analysis

Abstract
Background/Aims: Unlike the gene-poor Y chromosome, the X chromosome contains over 1,000 genes that are essential for viability of cells. Females have 2 X chromosomes, and thus female X-linked gene expression would be expected to be twice that of males. To adjust this imbalance, one of the 2 X-linked genes is often inactivated, and this is known as X-chromosome inactivation (XCI). However, recent studies described that a gene can be nonrandomly selected for inactivation from 2 X-linked genes and that XCI is not observed in some X-linked genes. Since this complex biological process has prevented efficient statistical association analyses, we propose a new statistical method against this uncertain biological process. ***Methods:*** The proposed method consists of 2 steps. First, *p* values for various biological processes are calculated and then combined into a single *p* value with the modified Fisher method and a minimum *p* value. ***Results:*** Our simulation results show that the proposed method is generally the most statistically efficient and is not sensitive to the unknown biological model. ***Conclusion:*** Therefore, we can conclude that the proposed approaches are robust against the various XCI processes for testing the association of X-linked single nucleotide polymorphisms with the disease of interest and the proposed method is a practical solution.

Introduction

X chromosomes are larger than Y chromosomes and carry over 1,000 genes that are essential for proper development and viability of the cell. As females have 2 X chromosomes unlike males, X-linked gene expression would be expected to be twice that of males, and thus gene expression in females needs to be adjusted to make it equivalent to males. This process is known as X-chromosome inactivation (XCI), which prevents females from having twice as many gene expression products as males. The XCI process was first described by Ohno et al. [1] in 1959. They showed that 1 X chromosome appeared similar to the autosomes but the other one was condensed and heterochromatic. Thereafter, Lyon [2] proposed that random inactivation of 1 female X chromosome explains the mottled phenotype of female mice that were heterozygous for coat color genes. However, recent studies have

Sungho Won
Department of Public Health Science
Seoul National University
1 Kwanak-ro, Kwanak-gu, Seoul 151-742 (Korea)
E-Mail won1@snu.ac.kr

Taesung Park
Department of Statistics
Seoul National University
1 Kwanak-ro, Kwanak-gu, Seoul 151-742 (Korea)
E-Mail tspark@stats.snu.ac.kr

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

also described a variety of XCI processes, such as nonrandom XCI and escape from XCI. The former implies that inactivation can be nonrandomly selected between 2 X-linked genes, and the latter implies that XCI is not observed in some X-linked genes [3–11]. This complex biological process has prevented efficient statistical association analyses, and it may partially explain the relatively small finding of significantly associated X-linked variants.

Multiple approaches were proposed for analysis of X-linked variants. Clayton [12] suggested 2 χ^2 tests with 1 and 2 degrees of freedom (dfs) tests. The χ^2 test with 1 df is used in a 2×2 table when the allele frequency in males and females can be assumed to be equal, and the χ^2 test with 2 dfs is the same as a conventional χ^2 test for association in the 3×2 contingency table. He assumed that the effect of males' homogeneous genotypes on phenotypes is equivalent to that of females' homozygous genotypes and thus coded females' genotypes as 0, 1, or 2, and males' genotypes as 0 or 2 [12]. However, even though this method is expected to be the most powerful when the underlying biological model is random XCI, it can lead to substantial power loss if the underlying biological models are nonrandom XCI or escaped XCI [13].

The highest statistical power can be achieved if the coding strategies reveal the expected differences [14], and Wang et al. [13] suggested a new statistical approach for various XCI processes, such as random XCI, nonrandom XCI, or escaped XCI. They used 0 or 1(2) for males' genotypes, and 0, *d*, or 2 for females' genotypes. Here, *d* is for heterogeneous genotypes in females and can be chosen between 0 and 2 depending on the XCI process. *d* should be lower than 1 for a nonrandom XCI toward the normal allele, and higher than 1 if a nonrandom XCI toward the deleterious allele is expected. Genotype coding of males depends on females' genotype. For random XCI or nonrandom XCI, males' genotypes are coded by 0 or 2, and for escaped XCI, males' genotypes are coded by 0 or 1. However, the mode of XCI is often unknown, and thus the most efficient choice of *d* is unclear. Wang et al. [13] suggested the minimum *p* value method among Wald statistics for logistic regression with various choices of *d* and conducted permutations for statistical inference. However, their approach is computationally very intensive and genome-wide search is time-consuming at the genome-wide significance level.

In this report, we extended the Cochran-Armitage test by Clayton [12] to handle the various modes of XCI. The proposed method computes the asymptotic distribution-based *p* values, and computational efficiency enables its application at the genome-wide level. We conducted extensive simulation studies and found that our proposed methods preserved reasonable statistical power for all XCI models even though it is not always best. Especially when the mode of XCI process occurs toward the normal allele, it always performs better than existing methods, and their gaps become bigger if samples with females larger than males are analyzed. Furthermore, statistical powers of the proposed methods are robust to various modes of XCI. The proposed methods were applied to type 2 diabetes (T2D) data, and identification of some promising single nucleotide polymorphisms (SNPs) revealed its practical value.

Methods

Notations and Disease Model

We assumed that there are N_m males and N_f females, with $N = N_m + N_f$. In this report, we considered only X-linked variants, and disease and normal alleles were assumed to be *A* and *a*, respectively. We denoted allele frequencies of females by q_A and q_a in the population. Allele frequencies of affected and unaffected female subjects were defined as q_a^a/q_A^a and q_a^u/q_A^u, respectively. The observed genotypes of females were designated *aa*, *Aa*, or *AA*, and in the population, the genotype frequencies of females were defined as p_{aa}, p_{Aa}, and p_{AA}. Genotype frequencies for affected and unaffected females were denoted by $p_{aa}^a/p_{Aa}^a/p_{AA}^a$ and $p_{aa}^u/p_{Aa}^u/p_{AA}^u$, respectively. Genotype frequencies of males in the population, and affected and unaffected males were denoted by p_a/p_A, p_a^a/p_A^a and p_a^u/p_A^u, respectively. It should be noted that genotype/allele frequencies of affected subjects are actually equal to the genotype/allele frequencies of unaffected subjects under the null hypothesis; that is, no association between disease status and genotypes. Therefore, the null hypothesis can be expressed by

$$H_0: p_{AA}^a = p_{AA}^u,\ p_{Aa}^a = p_{Aa}^u, \text{and } p_A^a = p_A^u.$$

If we assume that disease allele frequencies for males and females are the same and the Hardy-Weinberg equilibrium (HWE) is preserved, the null hypothesis can be simplified to

$$H_0: p_A^a = p_A^u = q_A^a = q_A^u.$$

Cochran-Armitage Trend Test for the Genotype Table

Statistical power for the Cochran-Armitage trend test is substantially affected by the coding strategy of each genotype [14], and we considered the different scores for genotypes depending on the mode of XCI. Our coding strategy was inspired by Wang et al. [13] but while their approach is based on permutation for logistic regression which is computationally very intensive, we consider asymptotic distribution-based inference for the Cochran-Armitage trend test. We assumed that males' genotypes are coded by m_a and m_A, and females' genotypes are by f_{aa}, f_{Aa}, and f_{AA}. For escaped XCI, 0 and 1 were assigned to m_a and m_A, respectively, and otherwise we use 0 and 2 for m_a and m_A, respectively. We proposed that $f_{aa} = 0$, $f_{Aa} = d$, and $f_{AA} = 2$. *d* is related to the skewness level of XCI. If disease alleles of heterozygous genotypes are less activated than normal alleles, *d* should be lower than 1; otherwise, *d* should be

Table 1. Notations

Gender	Genotypes	Number of cases	Number of controls	Coded values	Total
Female	*aa*	r_{aa}	s_{aa}	f_{aa}	n_{aa}
	Aa	r_{Aa}	s_{Aa}	f_{Aa}	n_{Aa}
	AA	r_{AA}	s_{AA}	f_{AA}	n_{AA}
Subtotal		R_f	S_f		N_f
Male	*a*	r_a	s_a	m_a	n_a
	A	r_A	s_A	m_A	n_A
Subtotal		R_m	S_m		N_m

larger than 1. If the amount of skewness varies among subjects, the statistical power may be maximized when the average level of skewed XCIs among subjects is used as d. Note that $d = 1$ indicates that randomly selected alleles from heterozygous genotypes of females are expressed. Furthermore, we assigned R_f/R_m and S_f/S_m as the numbers of affected females/males and unaffected females/males, respectively. Numbers of affected males with the genotypes *a* and *A* are denoted by r_a and r_A, and those of affected females are denoted by r_{aa}, r_{Aa}, and r_{AA}. Numbers of unaffected males with the genotypes *a* and *A* are denoted by s_a and s_A, and those of unaffected females are denoted by s_{aa}, s_{Aa}, and s_{AA}. Detailed notations about the number of genotypes for males and females are summarized in Table 1.

The choices of $f_{aa}/f_{Aa}/f_{AA}$ and m_a/m_A affect the statistical power and we consider L different coding strategies for XCI. For the coding strategy l, we can define scores for males and females, U_m^l and U_f^l, respectively, by

$$U_m^l = m_a^l\,(S_m r_a - R_m s_a) + m_A^l\,(S_m r_A - R_m s_A),$$

$$U_f^l = f_{aa}^l(S_f r_{aa} - R_f s_{aa}) + f_{Aa}^l(S_f r_{Aa} - R_f s_{Aa}) + f_{AA}^l(S_f r_{AA} - R_f s_{AA}).$$

Subsequently, under H_0, their expectations are equal to 0 as follows:

$$\begin{aligned} E\left(U_m^l\right) &= m_a^l\left(S_m E\left(r_a\right) - R_m E\left(s_a\right)\right) + m_A^l\left(S_m E\left(r_A\right) - R_m E\left(s_A\right)\right) \\ &= m_a^l\left(S_m R_m p_a - R_m S_m p_a\right) + m_A^l\left(S_m R_m p_a - R_m S_m p_a\right) = 0, \end{aligned}$$

$$\begin{aligned} E\left(U_f^l\right) &= f_{aa}^l\left(S_f E\left(r_{aa}\right) - R_f E\left(s_{aa}\right)\right) + f_{Aa}^l\left(S_f E\left(r_{Aa}\right) - R_f E\left(s_{Aa}\right)\right) \\ &\quad + f_{AA}^l\left(S_f E\left(r_{AA}\right) - R_f E\left(s_{AA}\right)\right) = f_{aa}^l\left(S_f R_f p_{aa} - R_f S_f p_{aa}\right) \\ &\quad + f_{Aa}^l\left(S_f R_f p_{Aa} - R_f S_f p_{Aa}\right) + f_{AA}^l\left(S_f R_f p_{AA} - R_f S_f p_{AA}\right) = 0. \end{aligned}$$

Variances of U_m^l and U_f^l can be calculated by

$$\begin{aligned} \mathrm{var}\left(U_m^l\right) &= m_a^{l\,2}\left(S_m{}^2\mathrm{var}\left(r_a\right) + R_m{}^2\mathrm{var}\left(s_a\right)\right) + m_A^{l\,2}\left(S_m{}^2\mathrm{var}\left(r_A\right) + R_m{}^2\mathrm{var}\left(s_A\right)\right) \\ &= m_a^{l\,2}\left(S_m{}^2 R_m p_a\left(1-p_a\right) + R_m{}^2 S_m p_a\left(1-p_a\right)\right) + m_A^{l\,2}\left(\begin{array}{l} S_m{}^2 R_m p_A\left(1-p_A\right) \\ + R_m{}^2 S_m p_A\left(1-p_A\right)\end{array}\right) \\ &= m_a^{l\,2}\left(S_m R_m N_m p_a(1-p_a)\right) + m_A^{l\,2}\left(S_m R_m N_m p_A(1-p_A)\right) \\ &= S_m R_m N_m\left\{m_a^{l\,2} p_a\left(1-p_a\right) + m_A^{l\,2} p_A\left(1-p_A\right)\right\}, \end{aligned}$$

$$\begin{aligned} \mathrm{var}\left(U_f^l\right) &= f_{aa}^{l\,2}\left(S_f{}^2\mathrm{var}\left(r_{aa}\right) + R_f{}^2\mathrm{var}\left(s_{aa}\right)\right) + \\ &\quad f_{Aa}^{l\,2}\left(S_f{}^2\mathrm{var}\left(r_{Aa}\right) + R_f{}^2\mathrm{var}\left(s_{Aa}\right)\right) + \\ &\quad f_{AA}^{l\,2}\left(S_f{}^2\mathrm{var}\left(r_{AA}\right) + R_f{}^2\mathrm{var}\left(s_{AA}\right)\right) \\ &= f_{aa}^{l\,2}\left(S_f{}^2 R_f p_{aa}\left(1-p_{aa}\right) + R_f{}^2 S_f p_{aa}\left(1-p_{aa}\right)\right) + \\ &\quad f_{Aa}^{l\,2}\left(S_f{}^2 R_f p_{Aa}\left(1-p_{Aa}\right) + R_f{}^2 S_f p_{Aa}\left(1-p_{Aa}\right)\right) + \\ &\quad f_{AA}^{l\,2}\left(S_f{}^2 R_f p_{AA}\left(1-p_{AA}\right) + R_f{}^2 S_f p_{AA}\left(1-p_{AA}\right)\right) \\ &= f_{aa}^{l\,2}\left(S_f R_f N_f p_{aa}(1-p_{aa})\right) + f_{Aa}^{l\,2}\left(S_f R_f N_f p_{Aa}(1-p_{Aa})\right) + \\ &\quad f_{AA}^{l\,2}\left(S_f R_f N_f p_{AA}(1-p_{AA})\right) \\ &= S_f R_f N_f\left\{\begin{array}{l} f_{aa}^{l\,2} p_{aa}\left(1-p_{aa}\right) + f_{Aa}^{l\,2} p_{Aa}\left(1-p_{Aa}\right) + \\ f_{AA}^{l\,2} p_{AA}\left(1-p_{AA}\right)\end{array}\right\}. \end{aligned}$$

Variances are functions of genotype frequencies, and under H_0, genotype frequencies in females can be estimated by

$$\hat{p}_{AA} = n_{AA}/N_f,\ \hat{p}_{Aa} = n_{Aa}/N_f$$

and under HWE,

$$\hat{p}_{AA} = \left(\frac{n_{AA}}{N_f} + \frac{0.5n_{Aa}}{N_f}\right)^2,\ \hat{p}_{Aa} = 2\left(\frac{n_{AA}}{N_f} + \frac{0.5n_{Aa}}{N_f}\right)\left(\frac{n_{aa}}{N_f} + \frac{0.5n_{Aa}}{N_f}\right).$$

For males, genotype frequencies are estimated with

$$\hat{p}_A = n_A/N_m,\ \hat{p}_a = n_a/N_m$$

If allele frequencies are expected to be the same between males and females, better estimators under HWE are

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa} + n_A}{2N_f + N_m},\ \hat{p}_{AA} = (\hat{p}_A)^2,\ \hat{p}_{Aa} = 2\hat{p}_A\left(1 - 2\hat{p}_A\right).$$

Based on these estimators, statistical associations of X-linked genes for a given choice of $(m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l$, and $f_{AA}^l)$ can be detected by the following test statistic:

$$T_l = \frac{U_m^l + U_f^l}{\sqrt{\mathrm{var}\left(U_m^l\right) + \mathrm{var}\left(U_f^l\right)}} \sim N\left(0,1\right) \text{ under } H_0.$$

Proposed Robust Statistics

T_l is a function of (m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l), and the most efficient choice can be determined if the mode of XCI is known. However, it is usually unknown, and alternatively statistics for L different choices of (m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l) can be combined. The robust statistics can be obtained by using the minimum p values or combining them with the modified Fisher methods. Statistics for different modes of XCI are correlated, and both statistics require the estimates for correlations among them. If we define the variance covariance matrix Φ by

$$\Phi = \begin{bmatrix} \mathrm{var}(T_1) & \cdots & \mathrm{cov}(T_1,T_L) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(T_L,T_1) & \cdots & \mathrm{var}(T_L) \end{bmatrix}, \text{ and } \sigma_{T_l} = \sqrt{\left\{\mathrm{var}\left(U_m^l\right) + \mathrm{var}\left(U_f^l\right)\right\}},$$

cov (T_l, T_l*) can be estimated by

$$\begin{aligned}
\mathrm{cov}\left(T_l,T_{l^*}\right) &= \frac{1}{\sigma_{T_l}\sigma_{T_{l^*}}}\mathrm{cov}\left(U_m^l + U_m^{l^*},\, U_f^l + U_f^{l^*}\right) \\
&= \frac{1}{\sigma_{T_l}\sigma_{T_{l^*}}}\left\{\mathrm{cov}\left(U_m^l,U_m^{l^*}\right) + \mathrm{cov}\left(U_f^l,U_f^{l^*}\right)\right\} \\
&= \frac{1}{\sigma_{T_l}\sigma_{T_{l^*}}}\left\{E\left(U_m^l,U_m^{l^*}\right) + E\left(U_f^l,U_f^{l^*}\right)\right\} \\
&= \frac{1}{\sigma_{T_l}\sigma_{T_{l^*}}}\left[S_m R_m N_m \begin{Bmatrix} m_a^l m_a^{l^*} p_a\left(1-p_a\right) + \\ m_A^l m_A^{l^*} p_A\left(1-p_A\right) \end{Bmatrix} + \right. \\
&\quad \left. S_f R_f N_f \begin{Bmatrix} f_{aa}^l f_{aa}^{l^*} p_{aa}\left(1-p_{aa}\right) + f_{Aa}^l f_{Aa}^{l^*} p_{Aa}\left(1-p_{Aa}\right) \\ + f_{AA}^l f_{AA}^{l^*} p_{AA}\left(1-p_{AA}\right) \end{Bmatrix}\right].
\end{aligned}$$

Based on these results, p values for T_l can be combined with a modified Fisher combining p value methods [15]. If p_l is the p value of T_l, we assume that

$$\chi^2 = -\sum_{l=1}^{L} 2\log p_l \sim c\cdot\chi^2\left(\mathrm{df} = f\right) \text{ under } H_0.$$

Brown [15] empirically showed that if T_l follows the multivariate normal distribution and $cor(T_l, T_{l^*}) = \rho_{ll^*}$, the covariance between $-2\log p_l$ and $-2\log p_{l^*}$ are approximately equal to

$$\mathrm{cov}\left(-2\log p_l, -2\log p_{l^*}\right) \approx \begin{cases} \rho_{ll^*}(3.25 + 0.75\rho_{ll^*}), & 0 \le \rho_{ll^*} \le 1 \\ \rho_{ll^*}(3.27 + 0.71\rho_{ll^*}), & -0.5 \le \rho_{ll^*} \le 0. \end{cases}$$

Thus under the null distribution, we have

$$\mathrm{E}\left(c\cdot\chi^2\left(\mathrm{df}=f\right)\right) = 2cf = E\left(-\sum_{l=1}^{L} 2\log p_l\right) = 2L$$

and

$$\begin{aligned}
&\mathrm{var}\left(c\cdot\chi^2\left(\mathrm{df}=f\right)\right) = 2c^2 f \\
&= \mathrm{var}\left(-\sum_{l=1}^{L} 2\log p_l\right) \sum_l \sum_{l^*} \mathrm{cov}\left(-2\log p_l,\, -2\log p_{l^*}\right) \\
&= \sum_l \mathrm{var}\left(-2\log p_l\right) + 2\sum_{l<l^*} \sum \mathrm{cov}\left(-2\log p_l,\, -2\log p_{l^*}\right) \\
&= 4\sum_{l=1}^{L} (\Phi)_{l,l} + 8\sum_{l=1}^{L-1}\sum_{l^*=l+1}^{L} (\Phi)_{l,l^*}.
\end{aligned}$$

Thus, by solving those equations, we can determine f and c as follows:

$$f = \frac{2L}{4\sum_{l=1}^{L}(\Phi)_{l,l} + 8\sum_{l=1}^{L-1}\sum_{l*=l+1}^{L}(\Phi)_{l,l^*}}$$

and

$$c = \frac{4\sum_{l=1}^{L}(\Phi)_{l,l} + 8\sum_{l=1}^{L-1}\sum_{l*=l+1}^{L}(\Phi)_{l,l^*}}{4L}.$$

p value for

$$-\sum_{l=1}^{L} 2\log p_l$$

can be estimated by using the estimated c and f, and will be denoted by P_{FIS} therein.

Alternatively, the minimum p value can be used as a robust test statistic. If we set T_{max} = max ($|T_1|$, …, $|T_L|$) and its realization by t_{max}, the minimum p value of the test statistic can be obtained by

$$\begin{aligned}
P_{MIN} &= P\{T_{max} > t_{max}\} = 1 - P\{T_{max} > t_{max}\} \\
&= 1 - P\{|T_1| < t_{max},\, |T_2| < t_{max},\, \ldots,\, |T_L| < t_{max}\}.
\end{aligned}$$

T_1, …, T_L are not independent, and their correlations were incorporated to their multivariate normal distribution. Then it was used to calculate $P\{|T_1| < t_{max}, |T_2| < t_{max}, \ldots, |T_L| < t_{max}\}$. Finally, for both P_{MIN} and P_{FIS}, we found that the proposed methods usually have good performance when $L = 6$, and our choices for (m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l) are shown in Table 2.

Simulation Studies

In our simulations, we assumed that there were 1,000 cases and 1,000 controls, and the number of males (females) in cases and controls were assumed to be the same. N_m:N_f were assumed to be 1:1, 1:2, and 2:1. Disease allele frequencies for (males, females) were assumed to be (0.2, 0.2), (0.3, 0.2), or (0.2, 0.3), and genotypes were generated under the HWE. It is known that levels of skewness of XCI can be locus-dependent and/or subject-dependent. Thus, we considered both scenarios in our simulations [7, 17, 19]. First, we assumed the same mode of XCI for all subjects at the considered X-linked variants and, second, the mode of XCI for each subject was randomly chosen from XCI_E, XCI_1, $XCI_{0.2}$, or $XCI_{1.8}$ with the same probability. Disease status for each individual was determined by the liability threshold model. If we denote the liability score and the coded genotype of subject i by y_i and X_i, respectively, y_i can be defined by summing the main generic effect, $X_i\beta$, and random error, ε_i, as follows:

$$y_i = \beta_0 + X_i\beta + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2).$$

β_0 was assumed to be 0. σ^2 indicates variances for random errors, and it was assumed to be 2. If we set

$$h_a^2 = \frac{2p_A\left(1-p_A\right)\beta^2}{2p_A\left(1-p_A\right)\beta^2 + \sigma^2},$$

h_a^2 indicates the relative proportion of variance explained by the disease susceptibility locus, which was assumed to be 0 for the type 1 error evaluations and 0.002 for statistical power evaluations. $h_a^2 = 0$ leads to $\beta = 0$, and $h_a^2 = 0.002$ does to $\beta = 0.1119$ if $p_A = 0.2$ and $\beta = 0.0977$ if $p_A = 0.3$. We considered the several modes of XCI by assigning m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l to X_i according to the gender and genotypes. Once the underlying liability scores of subjects

Table 2. Choices of $(m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, f_{AA}^l)$ for the various XCI processes

Model	Scenarios	m_a^l, m_A^l	$f_{aa}^l, f_{Aa}^l, f_{AA}^l$
1	Escaped XCI	0, 1	0, 1, 2
2	Skewed XCI with $d = 0.2$	0, 2	0, 0.2, 2
3	Skewed XCI with $d = 0.5$	0, 2	0, 0.5, 2
4	Random XCI with $d = 1$	0, 2	0, 1, 2
5	Skewed XCI with $d = 1.5$	0, 2	0, 1.5, 2
6	Skewed XCI with $d = 1.8$	0, 2	0, 1.8, 2

Table 3. Empirical type 1 error estimates

N_m:N_f	α	Empirical type 1 error estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
1:1	0.1	0.081 (0.069–0.092)	0.089 (0.077–0.101)	0.093 (0.080–0.106)	0.087 (0.075–0.099)	0.090 (0.077–0.103)
	0.05	0.039 (0.030–0.047)	0.051 (0.041–0.061)	0.044 (0.036–0.054)	0.049 (0.039–0.058)	0.052 (0.042–0.062)
	0.01	0.007 (0.003–0.011)	0.009 (0.005–0.013)	0.007 (0.003–0.011)	0.008 (0.004–0.012)	0.006 (0.003–0.009)
1:2	0.1	0.083 (0.070–0.095)	0.082 (0.070–0.094)	0.086 (0.074–0.098)	0.082 (0.070–0.094)	0.070 (0.059–0.081)
	0.05	0.041 (0.032–0.050)	0.045 (0.036–0.054)	0.045 (0.036–0.054)	0.047 (0.038–0.056)	0.036 (0.028–0.044)
	0.01	0.006 (0.003–0.009)	0.006 (0.003–0.009)	0.007 (0.003–0.011)	0.006 (0.003–0.009)	0.006 (0.003–0.009)
2:1	0.1	0.080 (0.068–0.092)	0.084 (0.072–0.096)	0.080 (0.068–0.092)	0.079 (0.067–0.091)	0.111 (0.097–0.125)
	0.05	0.032 (0.024–0.040)	0.033 (0.025–0.041)	0.044 (0.035–0.053)	0.030 (0.023–0.037)	0.052 (0.042–0.062)
	0.01	0.003 (0.001–0.005)	0.003 (0.001–0.005)	0.007 (0.003–0.011)	0.002 (0.001–0.003)	0.006 (0.003–0.009)

We assumed 1,000 cases and 1,000 controls. Empirical type 1 error estimates were estimated with 2,000 replicates.

were generated, they were transformed to disease status; subjects became affected if their liability scores were higher than a set threshold; otherwise, they were considered as unaffected. Prevalence was assumed to be 0.2, and threshold becomes 1.2362 if $p_A = 0.2$ and 0.8010 if $p_A = 0.3$.

Korea Associated Resource Cohort Data

We apply the proposed methods to a case-control association study of T2D with X-linked variants from Korea Associated Resource (KARE) cohorts that were recruited from South Korea. The KARE cohort was collected to construct an indicator of disease with genetic influences in an attempt to predict the occurrence of various diseases. There were 8,842 participants consisting of 4,183 males and 4,659 females, and they were recruited from Ansung and Ansan in the Gyeonggi Province of South Korea. Participants were between 40 and 69 years old. The 8,842 subjects were genotyped with the Affymetrix Genome-Wide Human SNP Array 5.0. We excluded SNPs with more than 5% missing genotype calls, very low minor allele frequencies lower than 0.05, and HWE p values were lower than 1.0×10^{-5} prior to analysis. As a result, 6,255 X-chromosome SNP markers passed the quality control filter, and they were used for genetic association analyses. Subjects with more than 5% missing genotype calls and incorrect sex information were excluded from the analyses.

Fasting plasma glucose and postprandial 2-h plasma glucose were used to diagnose T2D and measured every 2 years from 2001 to 2012. We used the definition of the American Diabetes Association to diagnose T2D (fasting plasma glucose ≥126 mg/dL; postprandial 2-h plasma glucose ≥200 mg/dL after a 75-g OGTT; and HbA1c ≥6.5%). During the follow-up period, 1,179 participants developed incident T2D and were classified as cases, i.e., 603 males and 576 females. Three controls for each case were matched based on age, sex, and the first 2 principal component scores were estimated with EIGENSTRAT [18]. Therefore, the proposed methods were applied to 603 male and 576 female T2D cases, and 1,809 male and 1,728 female controls for association analyses, and the remaining population substructure was adjusted by dividing each statistic by the genomic inflation factor.

Results

Method Evaluation with Simulated Data

To assess the performance of the proposed methods, we examined type 1 error rates and statistical powers under the 4 different modes of XCI. The most efficient choice of $(m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l$, and $f_{AA}^l)$ differs by XCI pro-

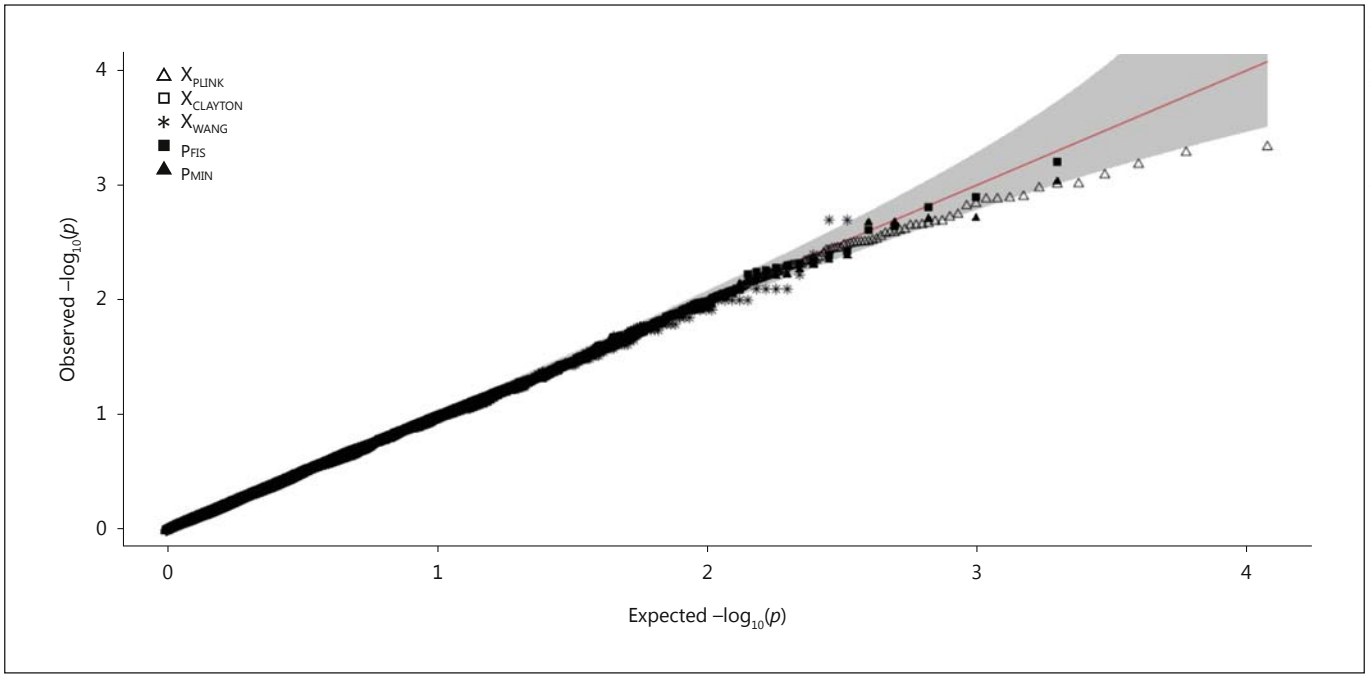

Fig. 1. Quantile-quantile plot of p values from association analyses of simulated data.

cesses, and if f_{AA}^{l} is coded as 2, the level of misspecification of XCI process is related with the choice of m_A^l and f_{Aa}^{l}. Thus, the relative performance of methods can be affected by the proportion of males, and we considered various ratios between numbers of males and females (N_m:N_f = 1:1, 1:2, or 2:1). Proportions of males in cases and controls were assumed to be the same. Clayton's [12] approach, the statistic implemented in PLINK [19], and the approach by Wang et al. [13] are denoted by $X_{CLAYTON}$, X_{PLINK}, and X_{WANG}, respectively. Permutation-based p values for X_{WANG} were obtained from 2,000 iterations.

We first evaluated statistical validity of the proposed methods with empirical type 1 error rates. Empirical type 1 error rates were estimated with 2,000 replicates. Table 3 shows that all methods preserve the 0.1, 0.05, and 0.01 nominal significance levels. Figure 1 also revealed that there was no inflation of the association analysis results. Second, statistical powers were evaluated with 2,000 replicates at the 0.05 significance level. Tables 4–6 show statistical power estimates for various modes of XCI. Table 4 shows results when the ratios of males and females are the same. Under XCI_E, an average of estimated powers of X_{PLINK} is 0.8, 2.0, 1.3, and 12.5% larger than those of P_{FIS}, P_{MIN}, $X_{CLAYTON}$, and X_{WANG}, respectively. X_{PLINK} assumes the escaped XCI, which explains the best performance of XCI_E. However, it is usually the least efficient for the other modes of XCI. The largest difference with the most efficient method was found for the $XCI_{0.2}$, which indicates the nonrandom XCI toward the normal allele. Under $XCI_{0.2}$, an average of estimated powers of P_{MIN} is the largest, and it is 2.6, 2.7, 10.2, and 13.9% larger than those of $X_{CLAYTON}$, P_{FIS}, X_{WANG}, and X_{PLINK}, respectively. Under XCI_1, an average of estimated powers of $X_{CLAYTON}$ is 0.3, 1.5, 5.5, and 9.7% larger than those of P_{FIS}, P_{MIN}, X_{PLINK}, and X_{WANG}, respectively. $X_{CLAYTON}$ uses (m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l) for the XCI_1, which explains its most efficiency. Under $XCI_{1.8}$, an average of estimated powers of P_{FIS} is the largest, and it is 0.2, 1.1, 2.0, and 8.3% larger than those of $X_{CLAYTON}$, P_{MIN}, X_{PLINK}, and X_{WANG}, respectively. Table 5 shows results for scenario where N_m:N_f = 1:2. Compared to the case where the numbers of males and females are the same, all statistics were more sensitive to modes of XCI. Furthermore, averages of statistical powers for all statistics are larger, which reveals that genetic association analyses of X-linked variants are less powerful than those of autosomal variants. Under XCI_E, an average of estimated powers of $X_{CLAYTON}$ is 0.1, 1.0, 1.2, and 10.4% larger than those of P_{FIS}, X_{PLINK}, P_{MIN}, and X_{WANG}, respectively. Under $XCI_{0.2}$, an average of estimated powers of P_{MIN} is 3.5,

Table 4. Empirical power estimates when N_m:N_f = 1:1

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	XCI_E	0.432 (0.410–0.454)	0.411 (0.389–0.433)	0.281 (0.261–0.301)	0.417 (0.395–0.439)	0.398 (0.377–0.419)
	$XCI_{0.2}$	0.434 (0.412–0.456)	0.551 (0.529–0.573)	0.466 (0.444–0.488)	0.549 (0.527–0.571)	0.574 (0.552–0.596)
	XCI_1	0.616 (0.595–0.637)	0.684 (0.664–0.704)	0.566 (0.544–0.588)	0.677 (0.657–0.697)	0.669 (0.648–0.690)
	$XCI_{1.8}$	0.793 (0.775–0.811)	0.795 (0.777–0.813)	0.707 (0.687–0.727)	0.797 (0.779–0.815)	0.787 (0.769–0.805)
0.3, 0.2	XCI_E	0.410 (0.388–0.432)	0.376 (0.355–0.392)	0.278 (0.258–0.298)	0.386 (0.365–0.407)	0.378 (0.357–0.399)
	$XCI_{0.2}$	0.467 (0.445–0.489)	0.561 (0.539–0.583)	0.477 (0.455–0.499)	0.562 (0.540–0.584)	0.581 (0.559–0.603)
	XCI_1	0.658 (0.637–0.679)	0.705 (0.685–0.725)	0.587 (0.565–0.609)	0.705 (0.685–0.725)	0.689 (0.669–0.709)
	$XCI_{1.8}$	0.789 (0.771–0.807)	0.786 (0.768–0.804)	0.713 (0.693–0.733)	0.792 (0.774–0.810)	0.784 (0.766–0.802)
0.2, 0.3	XCI_E	0.366 (0.345–0.387)	0.384 (0.363–0.405)	0.274 (0.254–0.294)	0.383 (0.362–0.404)	0.373 (0.352–0.394)
	$XCI_{0.2}$	0.438 (0.416–0.460)	0.565 (0.543–0.587)	0.507 (0.485–0.529)	0.564 (0.542–0.586)	0.601 (0.580–0.622)
	XCI_1	0.566 (0.544–0.588)	0.661 (0.640–0.682)	0.560 (0.538–0.582)	0.658 (0.637–0.679)	0.645 (0.624–0.666)
	$XCI_{1.8}$	0.695 (0.675–0.715)	0.749 (0.730–0.768)	0.668 (0.647–0.689)	0.748 (0.729–0.767)	0.732 (0.713–0.751)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

Table 5. Empirical power estimates when N_m:N_f = 1:2

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	XCI_E	0.435 (0.413–0.457)	0.447 (0.425–0.469)	0.333 (0.312–0.354)	0.446 (0.424–0.468)	0.438 (0.416–0.460)
	$XCI_{0.2}$	0.242 (0.223–0.261)	0.342 (0.321–0.363)	0.378 (0.357–0.399)	0.353 (0.332–0.374)	0.411 (0.389–0.433)
	XCI_1	0.495 (0.473–0.517)	0.594 (0.572–0.616)	0.521 (0.499–0.543)	0.587 (0.565–0.609)	0.571 (0.549–0.593)
	$XCI_{1.8}$	0.727 (0.707–0.747)	0.776 (0.758–0.794)	0.710 (0.690–0.730)	0.756 (0.737–0.775)	0.756 (0.737–0.775)
0.3, 0.2	XCI_E	0.424 (0.402–0.446)	0.441 (0.419–0.463)	0.318 (0.298–0.338)	0.441 (0.419–0.463)	0.424 (0.402–0.446)
	$XCI_{0.2}$	0.259 (0.240–0.278)	0.366 (0.345–0.387)	0.390 (0.369–0.411)	0.379 (0.358–0.400)	0.441 (0.419–0.463)
	XCI_1	0.507 (0.485–0.529)	0.582 (0.560–0.604)	0.492 (0.470–0.514)	0.575 (0.553–0.597)	0.561 (0.539–0.583)
	$XCI_{1.8}$	0.755 (0.736–0.774)	0.795 (0.777–0.813)	0.708 (0.688–0.728)	0.785 (0.767–0.803)	0.770 (0.752–0.788)
0.2, 0.3	XCI_E	0.389 (0.368–0.410)	0.391 (0.370–0.412)	0.315 (0.295–0.335)	0.388 (0.367–0.409)	0.380 (0.359–0.401)
	$XCI_{0.2}$	0.268 (0.249–0.287)	0.352 (0.331–0.373)	0.424 (0.402–0.446)	0.366 (0.345–0.387)	0.445 (0.423–0.467)
	XCI_1	0.412 (0.390–0.434)	0.497 (0.475–0.519)	0.507 (0.485–0.529)	0.490 (0.468–0.512)	0.506 (0.484–0.528)
	$XCI_{1.8}$	0.583 (0.561–0.605)	0.667 (0.646–0.688)	0.659 (0.638–0.680)	0.650 (0.629–0.671)	0.645 (0.624–0.666)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

6.6, 7.9, and 17.6% larger than those of X_{WANG}, P_{FIS}, $X_{CLAYTON}$, and X_{PLINK}, respectively. Under XCI_1, an average of estimated powers of $X_{CLAYTON}$ is the largest and it is 0.7, 1.2, 5.1, and 8.6% larger than those of P_{FIS}, P_{MIN}, X_{WANG}, and X_{PLINK}, respectively. Under $XCI_{1.8}$, an average of estimated powers of $X_{CLAYTON}$ is 1.6, 2.2, 5.4, and 5.8% larger than those of P_{FIS}, P_{MIN}, X_{WANG}, and X_{PLINK}, respectively. In Table 6, the number of males was assumed to be larger than that of females (N_m:N_f = 2:1). All statistics were not sensitive to modes of XCI processes as compared with the cases in which the number of females was larger than that of males. However, they were more

Table 6. Empirical power estimates when N_m:N_f = 2:1

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	XCI_E	0.382 (0.361–0.403)	0.334 (0.313–0.355)	0.266 (0.247–0.285)	0.342 (0.321–0.363)	0.350 (0.329–0.371)
	$XCI_{0.2}$	0.660 (0.639–0.681)	0.679 (0.659–0.699)	0.545 (0.523–0.567)	0.680 (0.660–0.700)	0.678 (0.658–0.698)
	XCI_1	0.720 (0.700–0.740)	0.713 (0.693–0.733)	0.611 (0.590–0.632)	0.714 (0.694–0.734)	0.710 (0.690–0.730)
	$XCI_{1.8}$	0.810 (0.793–0.827)	0.766 (0.747–0.785)	0.711 (0.691–0.731)	0.774 (0.756–0.792)	0.782 (0.764–0.800)
0.3, 0.2	XCI_E	0.393 (0.372–0.414)	0.341 (0.320–0.362)	0.283 (0.263–0.303)	0.356 (0.335–0.377)	0.368 (0.347–0.389)
	$XCI_{0.2}$	0.665 (0.644–0.686)	0.676 (0.655–0.697)	0.561 (0.539–0.583)	0.680 (0.660–0.700)	0.675 (0.654–0.696)
	XCI_1	0.722 (0.702–0.742)	0.707 (0.687–0.727)	0.611 (0.590–0.632)	0.710 (0.690–0.730)	0.715 (0.695–0.735)
	$XCI_{1.8}$	0.778 (0.760–0.796)	0.738 (0.719–0.757)	0.694 (0.674–0.714)	0.749 (0.730–0.768)	0.759 (0.740–0.778)
0.2, 0.3	XCI_E	0.382 (0.361–0.403)	0.346 (0.325–0.367)	0.274 (0.254–0.294)	0.358 (0.337–0.379)	0.359 (0.338–0.380)
	$XCI_{0.2}$	0.665 (0.644–0.686)	0.697 (0.677–0.717)	0.588 (0.566–0.610)	0.699 (0.679–0.719)	0.694 (0.674–0.714)
	XCI_1	0.716 (0.696–0.736)	0.718 (0.698–0.738)	0.610 (0.589–0.631)	0.720 (0.700–0.740)	0.713 (0.693–0.733)
	$XCI_{1.8}$	0.760 (0.741–0.779)	0.749 (0.730–0.768)	0.674 (0.653–0.695)	0.753 (0.734–0.772)	0.752 (0.733–0.771)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

Table 7. Empirical power estimates when mode of XCI process for each subject is randomly selected

N_m:N_f	*q*A (males, females)	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
1:1	0.2, 0.2	0.660 (0.639–0.681)	0.703 (0.683–0.723)	0.492 (0.470–0.514)	0.705 (0.685–0.725)	0.687 (0.667–0.707)
	0.3, 0.2	0.558 (0.536–0.580)	0.632 (0.611–0.653)	0.465 (0.443–0.487)	0.627 (0.606–0.648)	0.623 (0.602–0.644)
	0.2, 0.3	0.592 (0.570–0.614)	0.604 (0.583–0.625)	0.422 (0.400–0.444)	0.608 (0.587–0.629)	0.592 (0.570–0.614)
1:2	0.2, 0.2	0.506 (0.484–0.528)	0.596 (0.574–0.618)	0.455 (0.433–0.477)	0.571 (0.549–0.593)	0.640 (0.619–0.661)
	0.3, 0.2	0.427 (0.405–0.449)	0.523 (0.501–0.545)	0.419 (0.397–0.441)	0.500 (0.478–0.522)	0.589 (0.567–0.611)
	0.2, 0.3	0.498 (0.476–0.520)	0.562 (0.540–0.584)	0.419 (0.397–0.441)	0.538 (0.516–0.560)	0.602 (0.581–0.623)
2:1	0.2, 0.2	0.757 (0.738–0.776)	0.743 (0.724–0.762)	0.540 (0.518–0.562)	0.764 (0.745–0.783)	0.695 (0.675–0.715)
	0.3, 0.2	0.722 (0.702–0.742)	0.729 (0.710–0.748)	0.498 (0.476–0.520)	0.742 (0.723–0.761)	0.668 (0.647–0.689)
	0.2, 0.3	0.646 (0.625–0.667)	0.624 (0.603–0.645)	0.451 (0.429–0.473)	0.649 (0.628–0.670)	0.576 (0.545–0.589)

Mode of XCI process for each subject was randomly selected from XCI_E, $XCI_{0.2}$, XCI_1, and $XCI_{1.8}$, and statistical powers were estimated with 2,000 replicates at the 0.05 significance level.

sensitive, compared to the cases where the numbers of males and females were similar. Under XCI_E, an average of estimated powers of X_{PLINK} is 2.7, 3.4, 4.5, and 11.1% larger than those of P_{MIN}, P_{FIS}, $X_{CLAYTON}$, and X_{WANG}, respectively. Under $XCI_{0.2}$, an average of estimated powers of P_{FIS} is 0.2, 0.4, 2.3, and 12.2% larger than those of $X_{CLAYTON}$, P_{MIN}, X_{PLINK}, and X_{WANG}, respectively. Under XCI_1, an average of estimated powers of X_{PLINK} is the largest and it is 0.5, 0.7, 0.7, and 10.9% larger than those of P_{FIS}, P_{MIN}, $X_{CLAYTON}$, and X_{WANG}, respectively. Under $XCI_{1.8}$, an average of estimated powers of X_{PLINK} is 1.8, 2.4, 3.2, and 9.0% larger than those of P_{MIN}, P_{FIS}, $X_{CLAYTON}$, and X_{WANG}, respectively. The tables in the Appendix show statistical power estimates when d = 0.1, 0.4, 1.1, and 1.9, and N_m:N_f = 1:1, 1:2, and 1:1. General patterns are very similar as shown in Tables 4–6.

Table 7 shows the statistical powers when modes of XCI are different among subjects. The mode of XCI process for each subject was randomly selected from XCI_E, $XCI_{0.2}$, XCI_1, and $XCI_{1.8}$, and statistical powers were esti-

Table 8. Summary of simulation results

N_m:N_f	Mode of XCI					Method
	XCI_E	$XCI_{0.2}$	XCI_1	$XCI_{1.8}$	XCI_{MIXED}	
1:1	0.403	0.446	0.613	0.759	0.603	X_{PLINK}
	0.390	0.559	0.683	0.777	0.646	$X_{CLAYTON}$
	0.278	0.483	0.571	0.696	0.460	X_{WANG}
	0.395	0.558	0.680	0.779	0.647	P_{FIS}
	0.383	0.585	0.668	0.768	0.634	P_{MIN}
1:2	0.416	0.256	0.471	0.688	0.477	X_{PLINK}
	0.426	0.353	0.558	0.746	0.560	$X_{CLAYTON}$
	0.322	0.397	0.507	0.692	0.431	X_{WANG}
	0.425	0.366	0.551	0.730	0.536	P_{FIS}
	0.414	0.432	0.546	0.724	0.610	P_{MIN}
2:1	0.386	0.663	0.719	0.783	0.708	X_{PLINK}
	0.340	0.684	0.713	0.751	0.699	$X_{CLAYTON}$
	0.274	0.565	0.611	0.693	0.496	X_{WANG}
	0.352	0.686	0.715	0.759	0.718	P_{FIS}
	0.359	0.682	0.713	0.764	0.646	P_{MIN}

Averages of estimated powers for all simulation settings were calculated.

mated with 2,000 replicates at the 0.05 significance level. It will be denoted as XCI_{MIXED} in the remainder of this report. When numbers of males and females are the same, an average of estimated powers of P_{FIS} is 0.1, 1.3, 4.3, and 18.7% larger than those of $X_{CLAYTON}$, P_{MIN}, X_{PLINK}, and X_{WANG}, respectively. When females are larger than males (N_m:N_f = 1:2), the estimated powers of P_{MIN} are the largest, and they are followed by $X_{CLAYTON}$, and P_{FIS}. X_{WANG} and X_{PLINK} are the least efficient. An average of estimated powers of P_{MIN} is 5.0, 7.4, 13.3, and 17.9% larger than those of $X_{CLAYTON}$, P_{FIS}, X_{PLINK}, and X_{WANG}, respectively. When the number of males is larger than that of females (N_m:N_f = 2:1), P_{FIS} is usually the most efficient, and it is followed by X_{PLINK} and $X_{CLAYTON}$. An average of estimated powers of P_{FIS} is the largest and it is 1.0, 2.0, 7.2, and 22.2% larger than those of X_{PLINK}, $X_{CLAYTON}$, P_{MIN}, and X_{WANG}, respectively. The overall results from Tables 4 to 7 are summarized in Table 8.

In summary, we can conclude that P_{MIN} would be a reasonable choice for case-control association studies of X-linked variants with the nonrandom XCI toward the normal allele or if females are larger than males. P_{FIS} performs well if males are larger and XCI processes vary by subjects. Furthermore, P_{MIN} and P_{FIS} are usually less sensitive to the modes of XCI.

Application to KARE Cohort Data

We applied the proposed methods to the identification of X-chromosome-wide significant SNPs from KARE cohort study samples. Multiple testing problems were adjusted with a Bonferroni correction, and the Bonferroni-adjusted 0.05 significance level was 8.0×10^{-6}. Quantile-quantile plots in Figure 2 revealed that there is no inflation of association analyses. There was no X-chromosome-wide significant SNP, and it may be partially attributable to the insufficient sample size. Interestingly, $X_{CLAYTON}$ and P_{MIN} have very similar *p* values and it may be attributable to the similar numbers of males and females. The 10 most significant results are summarized in Table 9. These results suggest evidence for the association of X-linked variants with T2D. The most significant SNPs were rs11796450 and rs4898336, with *p* values lower than 1.0×10^{-4}. We also analyzed males and females separately. If males and females are separately analyzed, the sensitivity of statistics to modes of XCI becomes much smaller and we used the Cochran-Armitage trend test. Such stratified analyses are statistically powerful to detect the complicated interactions between sex and X-linked SNPs. *p* values for males and females are usually similar except for rs11796450, rs4898336, and rs5944901. Among the top 10 variants, rs11796450, rs5944902, rs5944901, rs4911827, rs17139421, and rs16978802 are located in the intergenic region on the X chromosome. rs4898336 and rs5986849 are located in a region that is upstream of the transcription elongation factor A like 6 (*TCEAL6*) and downstream of brain expressed X-linked 5 (*BEX5*), and rs2105854 is downstream of transcription elongation factor A like 2/6 (*TCEAL2*/*TCEAL6*). rs5971017 is located in

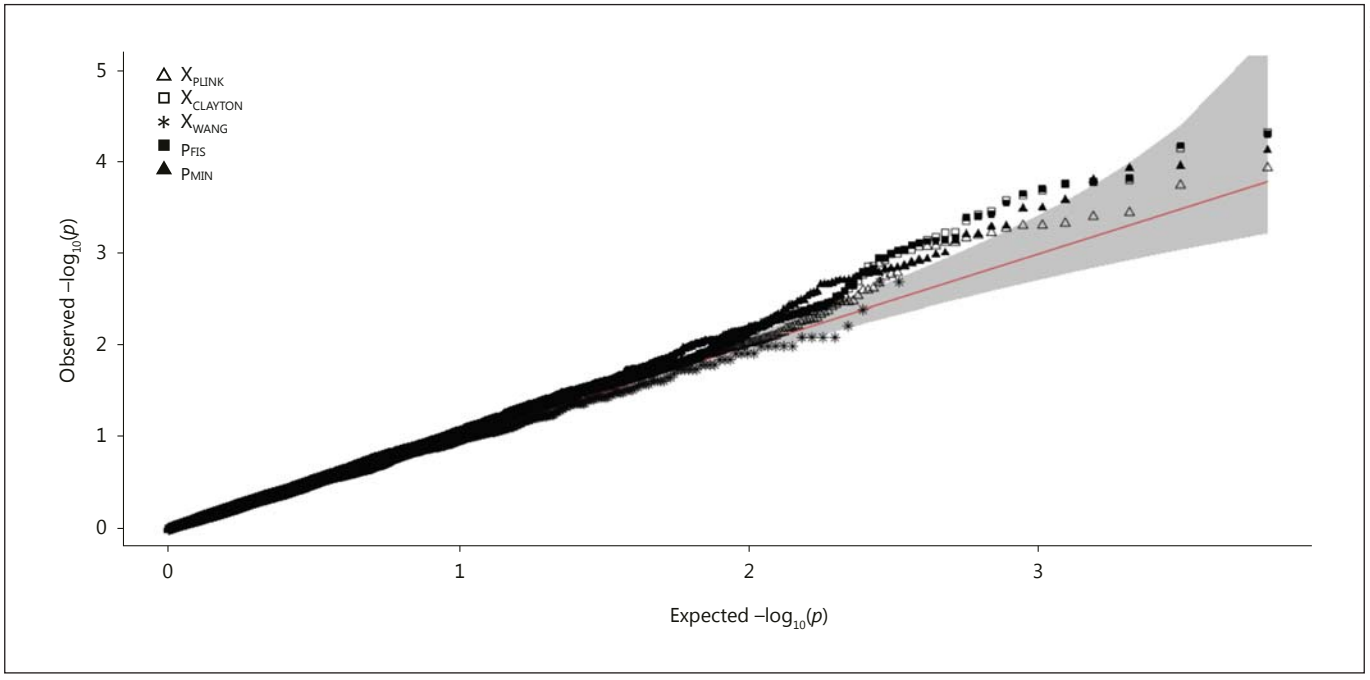

Fig. 2. Quantile-quantile plot of *p* values from association analyses of KARE data.

an intron of the Patched Domain Containing 1 (*PTCHD1*) gene locus on Xp22.11. It has been known that deletions in the 5′ flanking region of *PTCHD1* disrupted a complex noncoding RNA and potential regulatory elements. According to these results, P_{FIS} tends to have the most significant results, and it is followed by $P_{CLAYTON}$. Therefore, we can conclude that the proposed methods may be a reasonable choice for real data analyses.

Discussion

The biological process of X-linked variants is complex, and it may partially explain the relatively small number of X-linked disease susceptibility loci found with genome-wide association studies [3–11]. If the mode of XCI is known, such knowledge can be used to specify (m_a^l, m_A^l, f_{aa}^l, f_{Aa}^l, and f_{AA}^l). For instance, trimethyllysine hydroxylase epsilon is known to have a random XCI [7, 17]. However, modes of XCI process are usually unknown, and since there are several types of XCI in females such as random XCI, skewed XCI, and escaped XCI, how to consider such complex XCI processes in the analysis is not straightforward. Several methods for analysis of X-linked genes that consider the various XCI processes have been proposed. Although statistical methods that are suited for each process have been studied, it is very difficult to determine the most efficient statistical methods because the actual biological process is not known a priori. To overcome this drawback of the previous approach, we propose here a new association test that can account for various plausible biological models. The previously studied methods also have good control of the type 1 error, but the power is inferior to the proposed methods. Our simulation results show that our proposed methods are generally efficient and are not sensitive to the unknown biological model. Even though the true biological process for X-linked gene expression is often unknown, our extensive simulation studies revealed that the proposed approaches are usually efficient, and in particular, the power improvement is substantial with nonrandom XCI toward the normal allele and larger number of females than males. Furthermore, our proposed method is computationally very fast, and our genome-wide association analyses for T2D was completed within 243.31 CPU times with an Intel Xeon Processor E5-2620v2 ((6-core, 15 M Cache, 2.1 GHz, 7.2 GT/s, 80 W) × 2 ea). R-code for the proposed method can be freely downloaded from http://healthstat.snu.ac.kr/software/. Therefore, we can conclude that the proposed approaches are robust against the various XCI processes for testing the association of X-

Hum Hered 2016;82:50–63
DOI: 10.1159/000478048

Table 9. The most X-chromosome-wide significant 10 SNPs from GWAS with KARE data

SNP	BP	Major/minor alleles	MAFs	X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}	Only males	Only females	Genes
rs11796450	3983129	G/T	0.080	1.12×10^{-4}	4.81×10^{-5}	2.00×10^{-4}	4.82×10^{-5}	2.82×10^{-4}	5.53×10^{-4}	1.56×10^{-2}	Intergenic region on X chromosome
rs4898336	101401786	C/T	0.464	1.78×10^{-4}	6.95×10^{-5}	3.33×10^{-4}	6.44×10^{-5}	1.42×10^{-4}	7.97×10^{-4}	1.62×10^{-2}	Transcription elongation factor A like 6 (*TCEAL6*)/brain expressed X-linked 5 (*BEX5*)
rs5986849	101401535	C/T	0.465	3.84×10^{-4}	1.53×10^{-4}	9.00×10^{-4}	1.45×10^{-4}	2.37×10^{-4}	1.26×10^{-3}	2.38×10^{-2}	Transcription elongation factor A like 6 (*TCEAL6*)/brain expressed X-linked 5 (*BEX5*)
rs5944902	101204573	C/A	0.486	4.83×10^{-4}	1.71×10^{-4}	1.00×10^{-3}	1.61×10^{-4}	2.25×10^{-4}	1.05×10^{-2}	3.15×10^{-2}	Intergenic region on X chromosome
rs5971017	23114361	A/G	0.217	1.24×10^{-3}	1.65×10^{-4}	1.00×10^{-4}	1.65×10^{-4}	7.57×10^{-5}	1.63×10^{-4}	1.31×10^{-1}	Patched domain containing 1 (*PTCHD1*)
rs5944901	101204202	C/T	0.486	5.19×10^{-3}	2.00×10^{-4}	1.00×10^{-3}	1.89×10^{-4}	2.78×10^{-4}	1.36×10^{-2}	2.89×10^{-2}	Intergenic region on X chromosome
rs2105854	101389192	G/T	0.464	6.19×10^{-4}	2.24×10^{-4}	2.00×10^{-3}	2.14×10^{-4}	4.85×10^{-4}	1.27×10^{-2}	3.52×10^{-2}	Transcription elongation factor A like 2/6 (*TCEAL2*/*TCEAL6*)
rs4911827	114731229	T/G	0.247	7.34×10^{-4}	2.60×10^{-4}	5.00×10^{-4}	2.74×10^{-4}	4.81×10^{-4}	1.39×10^{-2}	3.66×10^{-2}	Intergenic region on X chromosome
rs17139421	3946340	C/T	0.063	6.64×10^{-4}	3.37×10^{-4}	1.00×10^{-3}	3.60×10^{-4}	5.45×10^{-4}	2.30×10^{-2}	3.02×10^{-2}	Intergenic region on X chromosome
rs16978802	3998041	A/G	0.103	7.36×10^{-4}	3.62×10^{-4}	3.00×10^{-3}	3.92×10^{-4}	6.10×10^{-4}	2.30×10^{-2}	3.31×10^{-2}	Intergenic region on X chromosome

linked SNPs with the disease of interest, and the proposed method is a practical solution.

However, in spite of the robustness of the proposed methods, there are some limitations. First, our proposed methods assume that all subject genotypes are independent. For association analyses, family-based samples are robust against the population substructures, and they are often utilized for candidate gene studies. Furthermore, subjects in the population-based samples can be correlated if there is population substructure and heritabilities are substantial. The proposed method can be simply extended to handle such family-based samples, in which the genetic correlation matrix is known as kinship coefficient matrix and can be simply calculated for X-linked variants, and to the quasi-likelihood score test [20]. Second, the proposed methods cannot adjust the effect of covariates. In genetic association analyses, age and gender are usually included as covariates. If effects of age and gender on disease status do not need to be adjusted, the proposed method may be useful in such scenarios. Third recent investigations showed that the amount of skewness of XCI can be affected by age [3, 5–7, 11, 21]. Furthermore, multiple X-linked variants which are in the same gene may have the same XCI process and it is better to apply the same XCI process to them. However, the proposed methods cannot be utilized in such scenarios and further extension which can handle heterogeneous modes of XCI by age and multiple SNPs are necessary. Last, if we can estimate the possible underlying mode of XCI, such information improves our understanding about diseases. P_{MIN} can provide some evidence about the underlying XCI process but any statistical inferences are not possible. In such a case, the Random Forest method [22] can be utilized. We are currently working on these issues and it will be further investigated with our future researches. Over the last decades, improvement in the genotyping/sequencing technology enabled genetic association analyses with tens of thousands of samples. However, even though X-linked genes are involved in many biological mechanisms of complex human diseases, significant results in the identification of X-linked variants have been rather limited. The novel finding of X-linked variants may be attributable to the lack of efficient statistical methods. Our report illustrates the importance of understanding these biological phenomena, and better understanding may be a key component for resolving the so-called missing heritability.

Acknowledgement

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (H I15C2165).

Appendix

Table A1. Empirical power estimates when N_m:N_f = 1:1

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	$XCI_{0.1}$	0.405 (0.383–0.427)	0.524 (0.502–0.546)	0.432 (0.410–0.454)	0.524 (0.502–0.546)	0.556 (0.534–0.578)
	$XCI_{0.4}$	0.476 (0.454–0.498)	0.590 (0.568–0.612)	0.488 (0.466–0.510)	0.588 (0.566–0.610)	0.595 (0.573–0.617)
	$XCI_{1.1}$	0.637 (0.616–0.658)	0.693 (0.673–0.713)	0.582 (0.560–0.604)	0.691 (0.671–0.711)	0.668 (0.647–0.689)
	$XCI_{1.9}$	0.805 (0.788–0.822)	0.806 (0.789–0.823)	0.690 (0.670–0.710)	0.806 (0.789–0.823)	0.807 (0.790–0.824)
0.3, 0.2	$XCI_{0.1}$	0.415 (0.393–0.437)	0.549 (0.527–0.571)	0.467 (0.445–0.489)	0.545 (0.523–0.567)	0.588 (0.566–0.610)
	$XCI_{0.4}$	0.459 (0.437–0.481)	0.603 (0.582–0.624)	0.499 (0.477–0.521)	0.560 (0.538–0.582)	0.606 (0.585–0.627)
	$XCI_{1.1}$	0.584 (0.562–0.606)	0.665 (0.644–0.686)	0.572 (0.550–0.594)	0.656 (0.635–0.677)	0.645 (0.624–0.666)
	$XCI_{1.9}$	0.696 (0.676–0.716)	0.757 (0.738–0.776)	0.682 (0.662–0.702)	0.752 (0.733–0.771)	0.747 (0.728–0.766)
0.2, 0.3	$XCI_{0.1}$	0.443 (0.421–0.465)	0.573 (0.551 0.595)	0.464 (0.442–0.486)	0.568 (0.546–0.590)	0.582 (0.560–0.604)
	$XCI_{0.4}$	0.495 (0.473–0.517)	0.592 (0.570–0.614)	0.489 (0.467–0.511)	0.589 (0.567–0.611)	0.584 (0.562–0.606)
	$XCI_{1.1}$	0.647 (0.626–0.668)	0.691 (0.671–0.711)	0.582 (0.560–0.604)	0.685 (0.665–0.705)	0.675 (0.654–0.696)
	$XCI_{1.9}$	0.818 (0.801–0.835)	0.803 (0.786–0.820)	0.743 (0.724–0.762)	0.807 (0.790–0.824)	0.805 (0.788–0.822)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

Table A2. Empirical power estimates when N_m:N_f = 1:2

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	$XCI_{0.1}$	0.213 (0.195–0.231)	0.305 (0.285–0.325)	0.223 (0.205–0.241)	0.325 (0.304–0.346)	0.395 (0.374–0.416)
	$XCI_{0.4}$	0.288 (0.268–0.308)	0.376 (0.355–0.397)	0.272 (0.252–0.292)	0.389 (0.368–0.410)	0.428 (0.406–0.450)
	$XCI_{1.1}$	0.524 (0.502–0.546)	0.603 (0.582–0.624)	0.495 (0.473–0.517)	0.597 (0.576–0.618)	0.574 (0.552–0.596)
	$XCI_{1.9}$	0.759 (0.740–0.778)	0.804 (0.787–0.821)	0.729 (0.710–0.748)	0.791 (0.773–0.809)	0.772 (0.754–0.790)
0.3, 0.2	$XCI_{0.1}$	0.224 (0.206–0.242)	0.319 (0.299–0.339)	0.243 (0.224–0.262)	0.339 (0.318–0.360)	0.414 (0.392–0.436)
	$XCI_{0.4}$	0.304 (0.284–0.324)	0.412 (0.390–0.434)	0.302 (0.282–0.322)	0.425 (0.403–0.447)	0.466 (0.444–0.488)
	$XCI_{1.1}$	0.537 (0.515–0.559)	0.605 (0.584–0.626)	0.521 (0.499–0.543)	0.602 (0.581–0.623)	0.586 (0.564–0.608)
	$XCI_{1.9}$	0.774 (0.756–0.792)	0.818 (0.801–0.835)	0.744 (0.725–0.763)	0.802 (0.785–0.819)	0.800 (0.782–0.818)
0.2, 0.3	$XCI_{0.1}$	0.241 (0.222–0.260)	0.329 (0.308–0.350)	0.264 (0.245–0.283)	0.347 (0.326–0.368)	0.443 (0.421–0.465)
	$XCI_{0.4}$	0.299 (0.279–0.319)	0.382 (0.361–0.403)	0.274 (0.254–0.294)	0.395 (0.374–0.416)	0.456 (0.434–0.478)
	$XCI_{1.1}$	0.443 (0.421–0.465)	0.535 (0.513–0.557)	0.442 (0.420–0.464)	0.529 (0.507–0.551)	0.516 (0.494–0.538)
	$XCI_{1.9}$	0.602 (0.581–0.623)	0.667 (0.646–0.688)	0.590 (0.568–0.612)	0.652 (0.631–0.673)	0.648 (0.627–0.669)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

Table A3. Empirical power estimates when N_m:N_f = 2:1

*q*A (males, females)	Mode of XCI	Empirical power estimates (95% CIs)				
		X_{PLINK}	$X_{CLAYTON}$	X_{WANG}	P_{FIS}	P_{MIN}
0.2, 0.2	$XCI_{0.1}$	0.642 (0.621–0.663)	0.673 (0.652–0.694)	0.568 (0.546–0.590)	0.669 (0.648–0.690)	0.672 (0.651–0.693)
	$XCI_{0.4}$	0.672 (0.651–0.693)	0.687 (0.667–0.707)	0.546 (0.524–0.568)	0.684 (0.664–0.704)	0.683 (0.663–0.703)
	$XCI_{1.1}$	0.744 (0.725–0.763)	0.733 (0.714–0.752)	0.667 (0.646–0.688)	0.734 (0.715–0.753)	0.734 (0.715–0.753)
	$XCI_{1.9}$	0.811 (0.794–0.828)	0.767 (0.748–0.786)	0.690 (0.670–0.710)	0.779 (0.761–0.797)	0.791 (0.773–0.809)
0.3, 0.2	$XCI_{0.1}$	0.646 (0.625–0.667)	0.668 (0.647–0.689)	0.598 (0.577–0.619)	0.667 (0.646–0.688)	0.667 (0.646–0.688)
	$XCI_{0.4}$	0.692 (0.672–0.712)	0.702 (0.682–0.722)	0.583 (0.561–0.605)	0.705 (0.685–0.725)	0.695 (0.675–0.715)
	$XCI_{1.1}$	0.740 (0.721–0.759)	0.715 (0.695–0.735)	0.598 (0.577–0.619)	0.721 (0.701–0.741)	0.725 (0.705–0.745)
	$XCI_{1.9}$	0.787 (0.769–0.805)	0.733 (0.714–0.752)	0.664 (0.643–0.685)	0.748 (0.729–0.767)	0.763 (0.744–0.782)
0.2, 0.3	$XCI_{0.1}$	0.657 (0.636–0.678)	0.694 (0.674–0.714)	0.611 (0.590–0.632)	0.692 (0.672–0.712)	0.696 (0.676–0.716)
	$XCI_{0.4}$	0.671 (0.650–0.692)	0.695 (0.675–0.715)	0.575 (0.553–0.597)	0.696 (0.676–0.716)	0.693 (0.673–0.713)
	$XCI_{1.1}$	0.725 (0.705–0.745)	0.729 (0.710–0.748)	0.612 (0.591–0.633)	0.732 (0.713–0.751)	0.728 (0.708–0.748)
	$XCI_{1.9}$	0.772 (0.754–0.790)	0.749 (0.730–0.768)	0.685 (0.665–0.705)	0.756 (0.737–0.775)	0.765 (0.737–0.775)

We considered different disease allele frequencies for males and females. Empirical powers were estimated with 2,000 replicates at the 0.05 significance level.

References

1 Ohno S, Kaplan WD, Kinosita R: Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. Exp Cell Res 1959;18:415–418.
2 Lyon MF: Gene action in the X-chromosome of the mouse (*Mus musculus* L.). Nature 1961; 190:372–373.
3 Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF: X-chromosome inactivation patterns of 1,005 phenotypically unaffected females. Am J Hum Genet 2006;79:493–499.
4 Belmont JW: Genetic control of X inactivation and processes leading to X-inactivation skewing. Am J Hum Genet 1996;58:1101.
5 Busque L, Paquette Y, Provost S, Roy DC, Levine RL, Mollica L, Gilliland DG: Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. Blood 2009;113:3472–3474.
6 Chagnon P, Provost S, Belisle C, Bolduc V, Gingras M, Busque L: Age-associated skewing of X-inactivation ratios of blood cells in normal females: a candidate-gene analysis approach. Exp Hematol 2005;33:1209–1214.
7 Minks J, Robinson WP, Brown CJ: A skewed view of X chromosome inactivation. J Clin Invest 2008;118:20–23.
8 Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF: Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. Am J Hum Genet 2002;71:168–173.
9 Struewing JP, Pineda MA, Sherman ME, Lissowska J, Brinton LA, Peplonska B, Bardin-Mikolajczak A, Garcia-Closas M: Skewed X chromosome inactivation and early-onset breast cancer. J Med Genet 2006;43:48–53.
10 Willard H: The sex chromosomes and X chromosome inactivation; in Sriver CR, Beaudet AL, Sly WS, Valle D (eds): The Metabolic and Molecular Bases of Inherited Disease. New York, McGraw-Hill, 1995, pp 719–737.
11 Wong CC, Caspi A, Williams B, Houts R, Craig IW, Mill J: A longitudinal twin study of skewed X chromosome-inactivation. PLoS One 2011;6:e17873.
12 Clayton D: Testing for association on the X chromosome. Biostatistics 2008;9:593–600.
13 Wang J, Yu R, Shete S: X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. Genet Epidemiol 2014;38: 483–493.
14 Agresti A, Kateri M: Categorical Data Analysis. Berlin, Springer, 2011.
15 Brown MB: 400: a method for combining non-independent, one-sided tests of significance. Biometrics 1975;31:987–992.
16 Cotton AM, Ge B, Light N, Adoue V, Pastinen T, Brown CJ: Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome Biol 2013;14:R122.
17 Carrel L, Willard HF: X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 2005;434:400–404.
18 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.
19 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.
20 Thornton T, McPeek MS: Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 2007;81:321–337.
21 Sharp A, Robinson D, Jacobs P: Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet 2000;107:343–349.
22 Winham SJ, Jenkins GD, Biernacka JM: Modeling X chromosome data using random forests: conquering sex bias. Genet Epidemiol 2016;40:123–132.