HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

Genet Epidemiol. 2016 September ; 40(6): 475–485. doi:10.1002/gepi.21979.

FARVATX: FAmily-based Rare Variant Association Test for X-linked genes

Sungkyoung Choi[1], **Sungyoung Lee**[1], **Dandi Qiao**[2], **Megan Hardin**[2,3], **Michael H. Cho**[2,3], **Edwin K Silverman**[2,3], **Taesung Park**[1,4,*], and **Sungho Won**[1,5,6,*]

[1]Interdisciplinary Program in bioinformatics, Seoul National University, Seoul, Korea

[2]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States of America

[3]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, United States of America

[4]Department of Statistics, Seoul National University, Seoul, Korea

[5]Department of Public Health Science, Seoul National University, Seoul, Korea

[6]Institute of Health and Environment, Seoul National University, Seoul, Korea.

Abstract

Although the X chromosome has many genes that are functionally related to human diseases, the complicated biological properties of the X chromosome have prevented efficient genetic association analyses, and only a few significantly associated X-linked variants have been reported for complex traits. For instance, dosage compensation of X-linked genes is often achieved via the inactivation of one allele in each X-linked variant in females; however, some X-linked variants can escape this X chromosome inactivation. Efficient genetic analyses cannot be conducted without prior knowledge about the gene expression process of X-linked variants, and misspecified information can lead to power loss. In this report, we propose new statistical methods for rare X-linked variant genetic association analysis of dichotomous phenotypes with family-based samples. The proposed methods are computationally efficient and can complete X-linked analyses within a few hours. Simulation studies demonstrate the statistical efficiency of the proposed methods, which were then applied to rare-variant association analysis of the X chromosome in chronic obstructive pulmonary disease (COPD). Some promising significant X-linked genes were identified, illustrating the practical importance of the proposed methods.

Keywords

X chromosome; X chromosome inactivation; extended families; rare variants; genetic association analysis

[*]**Correspondence to:** Sungho Won, Dept. of Public Health Science, Seoul National University, 1 Gwanak-ro Gwanak-gu, Seoul 151-742, Korea. won1@snu.ac.kr, Tel: +82-2-880-2714; Taesung Park, Department of Statistics, Seoul National University, 1 Gwanak-ro Gwanak-gu, Seoul 151-742, Korea. tspark@stats.snu.ac.kr, Tel: +82-2-880-8924.

Conflict of Interest: none declared.

Introduction

Due to the relatively large size of the X chromosome, many X-linked genes have important functions, and significant associations of several X-linked variants have been identified for diverse phenotypes, including blood pressure, hematological traits, obesity, HDL cholesterol, and Type-1 diabetes [Ahituv, et al. 2007; Auer, et al. 2014; Blakemore, et al. 2009; Cohen, et al. 2004; Gaukrodger, et al. 2005; Nejentsev, et al. 2009]. However, most successful results from genome-wide association studies have been from autosomes, and significant results for X-linked variants are relatively few in number. There are multiple potential reasons for this, but it is at least partially attributable to the complex biological properties of X-linked variants, which make efficient genetic association analyses more challenging. For instance, while females inherit X chromosomes from both parents, males inherit a single maternal X chromosome, and there is some empirical evidence that in females genes for some X-linked variants are expressed twice as highly as in males [Brown and Greally 2003; Carrel and Willard 2005; Shapiro, et al. 1979]. In contrast, dosage compensation for other X-linked variants can be achieved by the selection, and silencing of maternal or paternal genes via either random or nonrandom mechanisms [Lyon 1961]. Under nonrandom X chromosome inactivation (XCI), either the maternal or paternal genes are relatively more activated [Belmont 1996; Plenge, et al. 2002], and the amount of skewness is sometimes related to age or disease status [Amos-Landgraf, et al. 2006; Busque, et al. 1996; Chagnon, et al. 2005; Knudsen, et al. 2007; Minks, et al. 2008; Sharp, et al. 2000; Wong, et al. 2011]. However in spite of this knowledge about gene expression process of X-linked variants, there are very few statistical methods applicable to the complicated biological process of X-linked genes and thus development of new statistical methods is necessary.

Several statistical methods have been proposed for detecting statistically associated X-linked variants of phenotypes. In prospective analyses, genetic association can be simply detected by using only female subjects or by incorporating gender as a covariate, and thus in this report we focus on retrospective study designs. In retrospective analyses, genetic associations of autosomal variants are detected by comparing genetic distributions between affected and unaffected individuals, and the Cochran Armitage trend test, which compares minor allele frequencies (MAFs) in affected and unaffected individuals [Armitage 1955; Clayton 2008; Sasieni 1997; Zheng, et al. 2003; Zhu and Xiong 2012], is often utilized to assess the association. For X-linked variants, there is heterogeneity of genetic distributions between males and females, which is often handled by extending the Cochran-Armitage test for genetic association analyses of X-linked variants [Clayton 2008; Zheng, et al. 2007]. For family-based samples, the M_{QLS} method [Thornton and McPeek 2007] has also been modified for use in common X-linked variant association analysis, and Schaid et al. [Schaid, et al. 2013] have proposed new methods for rare X-linked variant association analysis with family-based samples. For instance, since the complex biological properties of the X chromosome affect statistical power, XCI has been addressed by using the same coded values for males with hemizygous disease genotypes as those used with females with homozygous disease genotypes [Clayton 2008; Thornton, et al. 2012]. However XCI can be nonrandom and is sometimes completely escaped, which leads to some power loss for these approaches. Therefore, in this report we propose a family-based rare variant association test

for X-linked genes (***FARVATX***) that is applicable to various biological models. Due to the nature of our statistic, the proposed method can also be applied to family-based designs with dichotomous phenotype, and we show with extensive simulation studies that the proposed methods perform better than the existing approaches. We applied the coding strategy that was suggested by Wang et al. [Wang, et al. 2014] in population-based design. The proposed methods were applied to an association analysis of families with chronic obstructive pulmonary disease (COPD). Some promising genes were identified with the proposed methods, thereby illustrating the practical value of these methods.

Methods

Notation

We assume that there are n families and n_i individuals in family i, and the total sample size is denoted by $N=\sum_{i=1}^{n} n_i$. We assume that genotypes for M rare variants on the X chromosome are available. We let x_{ij}^m be the coded genotype of an individual j in a family i for a variant m, with allowed values of 0, 1, or 2 for a female, and 0 or 1 for a male individual, depending on the number of minor alleles. We denote the disease prevalence by q and assume that y_{ij} is coded as 1 for affected individuals, q for individuals with missing phenotype, and 0 for unaffected individuals. In retrospective analyses, genetic association is detected by comparing genetic distributions of affected and unaffected individuals, and it has been shown that the statistical efficiency can be improved by modifying the phenotype [Lange and Laird 2002; Thornton and McPeek 2007]. We let μ_{ij} be the offset that is define by disease prevalence or the best linear unbiased predictor (BLUP) from the linear mixed model [Won and Lange 2013], and set $t_{ij} = y_{ij} - \mu_{ij}$. Then, if we represent the column vectors that comprise x_{ij}^m and t_{ij} for all individuals in a family i by $\mathbf{X}_i^m$ and $\mathbf{T}_i$ respectively, the genotype matrix and phenotype vector can be defined by

$$\mathbf{X}^m=\begin{pmatrix}\mathbf{X}_1^m\\ \vdots\\ \mathbf{X}_n^m\end{pmatrix},\mathbf{X}=\begin{pmatrix}\mathbf{X}^1 & \cdots & \mathbf{X}^M\end{pmatrix},\quad \text{and}\quad \mathbf{T}=\begin{pmatrix}\mathbf{T}_1\\ \vdots\\ \mathbf{T}_n\end{pmatrix}.$$

Variance covariance matrix

We assume that $\sigma_{mm'}$ is a covariance between x_{ij}^m and $x_{ij}^{m'}$ when an individual j in a family i is a male, and the genetic variance-covariance matrix between M markers in males is

$$\begin{pmatrix}\sigma_{11} & \cdots & \sigma_{1M}\\ \vdots & \ddots & \vdots\\ \sigma_{M1} & \cdots & \sigma_{MM}\end{pmatrix}$$

We assume that h_{ij} is an inbreeding coefficient for an individual j in a family i, and thus if an individual j is a male, h_{ij} becomes 0. $\pi_{ij,i'j'}$ is a kinship coefficient between an individual j in a family i and an individual j' in a family i'. It should be noted that $\pi_{ij,i'j'}$ is a function of

gender, and can be deductively calculated [Thornton, et al. 2012]. If i and i' are different, $\pi_{ij,i'j'}$ becomes 0. We consider individuals j and j' in a family i, and if an individual j is a descendant of j', $\pi_{ij,ij'}$ can be derived based on Table 1. We consider the case where individual j in a family i is not a descendant of j' in a large family. If we let $m(j')$ and $f(j')$ indicate the mother and father of j', respectively, $\pi_{ij,ij'}$ can be recursively calculated as follows:

1. $\pi_{ij,ij'} = \pi_{ij,im(j')}/2$, if j' is a male.
2. $\pi_{ij,ij'} = \pi_{ij,im(j')}/2 + \pi_{ij,if(j')}$, if j' is a female.

If we define $\mathbf{\Phi}$ by

$$\mathbf{\Phi}_{i,i}\begin{pmatrix} 1+h_{i1} & 2\pi_{i1,i2} & \cdots \\ 2\pi_{i2,i1} & 1+h_{i2} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}, \mathbf{\Phi}_{i,i'}=\begin{pmatrix} 0 & 0 & \cdots \\ 0 & 0 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}(i \neq i'), \quad \text{and} \quad \mathbf{\Phi}=\begin{pmatrix} \mathbf{\Phi}_{1,1} & \mathbf{\Phi}_{1,2} & \cdots \\ \mathbf{\Phi}_{2,1} & \mathbf{\Phi}_{2,2} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix},$$

then we have $\mathrm{var}\,(\mathbf{X}^m) = \sigma^2_{mm}\mathbf{\Phi}$.

If we let $\mathbf{l}_N$ be the $N\times 1$ column vector of which the elements are 1 for male and 2 for female, respectively, the best linear unbiased estimator for $E(\mathbf{X})$ under the null hypothesis can be derived, with some modification of the methods of McPeek et al [McPeek, et al. 2004], by

$$\hat{E}\,(\mathbf{X}) = \mathbf{l}_N\left(\mathbf{l}_N^t\,\mathbf{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\,\mathbf{\Phi}^{-1}\mathbf{X},$$

, and $\mathbf{\Sigma}$ can be estimated by

$$\hat{\mathbf{\Sigma}} = \frac{1}{(N-1)}\left[\mathbf{X}^t\mathbf{\Phi}^{-1}\mathbf{X} - \left(\mathbf{l}_N^t\,\mathbf{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\left(\mathbf{l}_N^t\,\mathbf{\Phi}^{-1}\mathbf{X}\right)^2\right].$$

Weighted quasi-likelihood score

We assume that $\mathbf{D}_d$ is a $N\times N$ diagonal matrix, and its diagonal elements are 1 or d if the corresponding individuals are males or females, respectively. X-linked gene expression processes are considered by replacing the genotype matrix $\mathbf{X}$ by $\mathbf{D}_d\mathbf{X}$. $\mathbf{D}_d\mathbf{X}$ will be called the weighted quasi-likelihood score in the remainder of this report. The efficient choice of d is related to the gene expression process and can be obtained by considering the relative proportion of each genotype's expression [Clayton 2008]. In particular, homozygous disease genotypes are not usually observed for rare variants; thus, an approximately efficient coding strategy can be chosen by comparing gene expression levels for heterozygous disease genotypes in females and hemizygous disease genotypes in males. Therefore under our coding strategy, XCI and escaped XCI (E-XCI) are efficiently tested with $d = 0.5$ and $d = 1$, respectively. We also have considered another simulation scenario for skewed XCI (S-XCI) owing to nonrandom XCI. S-XCI have been defined using an arbitrary threshold as inactivation of deleterious or normal allele in more than 75% cells [Abkowitz, et al. 1998].

We assumed that the value of d was set as 0.75 or 0.25 to represent S-XCI toward to the deleterious allele or the normal allele, respectively.

Rare X-linked variant association tests for XCI, E-XCI, and S-XCI

The quasi-likelihood-based score [Won and Lange 2013] for $\mathbf{D}_d\mathbf{X}$ can be defined by

$$\mathbf{T}^t\left(\mathbf{D}_d\mathbf{X}-E\left(\mathbf{D}_d\mathbf{X}\right)\right)=\mathbf{T}^t\mathbf{D}_d\left(\mathbf{X}-E\left(\mathbf{X}\right)\right).$$

Because $E(\mathbf{X})$ can be estimated by $\mathbf{l}_N\left(\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}\mathbf{X}$, the quasi-likelihood score becomes $\mathbf{T}^t\mathbf{D}_d\left(\mathbf{I}_N-\mathbf{l}_N\left(\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}\right)\mathbf{X}=\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}$ where $\mathbf{P}=\boldsymbol{\Phi}^{-1}-\boldsymbol{\Phi}^{-1}\mathbf{l}_N\left(\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\boldsymbol{\Phi}^{-1}$. If we let $\mathbf{H}=\boldsymbol{\Phi}-\mathbf{l}_\mathbf{N}\left(\mathbf{l}_\mathbf{N}^\mathbf{t}\boldsymbol{\Phi}^{-1}\mathbf{l}_\mathbf{N}\right)^{-1}\mathbf{l}_\mathbf{N}^\mathbf{t}$, we can simply show that

$$\mathrm{cov}\left(\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}^\mathbf{m},\mathbf{T}^\mathbf{t}\mathbf{D}_\mathbf{d'}\boldsymbol{\Phi}\mathbf{P}\mathbf{X}^\mathbf{m'}\right)=\left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\sigma_{mm'},$$

and thus we have

$$\mathrm{cov}\left(\left(\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\right)^t,\left(\mathbf{T}^t\mathbf{D}_{d'}\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\right)\right)=\left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\boldsymbol{\Sigma}.$$

It has been empirically shown that weighting each variant can be an efficient strategy to improve statistical power for rare variant association analyses [Madsen and Browning 2009]. We let the weight for variant m be w_m, and the diagonal matrix for which the diagonal element m is w_m be $\mathbf{W}$. If we let p_m be the MAF for a variant m, we used $Beta(p_m; 1, 25)$ as w_m. Then scores for burden [Li and Leal 2008] and variance component [Neale, et al. 2011; Wu, et al. 2011] tests can be respectively defined by

$$\frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\left(\mathbf{0}.\mathbf{I}_\mathbf{M}+(\mathbf{1}-\mathbf{0})\cdot\mathbf{1}_\mathbf{M}\mathbf{1}_\mathbf{M}^\mathbf{t}\right)\mathbf{W}\mathbf{X}^\mathbf{t}\mathbf{P}\boldsymbol{\Phi}\mathbf{D}_\mathbf{d}\mathbf{T},\quad\textbf{and}$$
$$\frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\left(\mathbf{1}.\mathbf{I}_\mathbf{M}+(\mathbf{1}-\mathbf{1})\cdot\mathbf{1}_\mathbf{M}\mathbf{1}_\mathbf{M}^\mathbf{t}\right)\mathbf{W}\mathbf{X}^\mathbf{t}\mathbf{P}\boldsymbol{\Phi}\mathbf{D}_\mathbf{d}\mathbf{T}.$$

These are extensions of FARVAT statistics [Choi, et al. 2014]. We let $\mathbf{R}_c=c\mathbf{I}_M+(1-c)\,\mathbf{1}_M\mathbf{1}_M^t$ and define

$$S_c^{(d)}=\frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\mathbf{R}_\mathbf{c}\mathbf{W}\mathbf{X}^\mathbf{t}\mathbf{P}\boldsymbol{\Phi}\mathbf{D}_\mathbf{d}\mathbf{T}.$$

We let p-values for $S_c^{(d)}$ be $P_c^{(d)}$, and denote $P_0^{(d)}$ and $P_1^{(d)}$ by ***FARVAT-XB*$_{(d)}$** and ***FARVAT-XC*$_{(d)}$**. It should be noted that the formal corresponds to the burden-type statistic and the latter does SKAT-type statistic. The SKAT-O-type statistic [Lee, et al. 2012b] can be defined by

$$\min\left\{P_0^{(d)}, P_{0,1}^{(d)}, \ldots, P_1^{(d)}\right\},$$

and we denote its *p*-values by ***FARVAT-XO*****$_{(d)}$**. *P*-values can be calculated by the numerical algorithms for ***FARVAT*** statistics [Choi, et al. 2014].

If the biological gene expression processes of X-linked genes are not clear, the proposed statistics may be sensitive to the choice of d, and a robust statistic needs to be provided. We calculate ***FARVAT-XB*****$_{(d)}$** or ***FARVAT-XC*****$_{(d)}$** for various choices of d, and then combine them to a single *p*-value by using extended Fisher's method for correlated p-values [Brown 1975]. We denote its *p*-value by ***FARVAT-XD*** where 0, 0.05, 0.1, …, 0.95, and 1 were considered for d_1, …, and d_L. The detailed algorithm is described in Supplementary Material.

Simulation Studies

To investigate the performance of the proposed methods, we performed simulation studies for various family structures (see Figure 1 for detailed information). We considered trios with a son or a daughter, and large families with 10 individuals that extended over three generations and had different numbers of males and females. MAFs were generated from a uniform distribution $U(0, 0.01)$, and genotype frequencies were calculated under Hardy-Weinberg Equilibrium (HWE). If we let p_m be the MAF for a variant m, founders' genotypes were generated with a binomial distribution $B(2, p_m)$, and offspring's genotypes were obtained by simulated Mendelian transmission, assuming no recombination. Phenotypes for each individual were generated with a liability threshold model, and liabilities were determined by summing the phenotypic mean $\left(\bar{\mu}\right)$, polygenic effect $\left(\sigma_g^2\right)$, common environmental effect $\left(\sigma_c^2\right)$, main genetic effect and random error $\left(\sigma_e^2\right)$. Random errors were independently generated from $N\left(0, \sigma_e^2{=}1/3\right)$. The polygenic effect for founders was independently generated from $N\left(0, \sigma_g^2{=}1/3\right)$, and for non-founders, averages of maternal and paternal polygenic effects were combined with values independently sampled from $N\left(0, 0.5\sigma_g^2\right)$. Common environmental effects were assumed to be the same for all individuals in each family and were generated from $N\left(0, \sigma_c^2{=}1/3\right)$. For main genetic effects, we assumed that there were M rare variants, and genetic effects for each rare variant were obtained by the product of β_m, the number of disease alleles, and d. If we let h_a^2 be the proportion of phenotypic variance explained by the main genotype, β_m were sampled from $U(1.0, v)$ and v was calculated by

$$v = \sqrt{\frac{\left(\sigma_g^2+\sigma_c^2+\sigma_e^2\right) h_a^2}{\left(1-h_a^2\right) d^2 \sum_{m=1}^{M} \beta_m^2 2p_m\left(1-p_m\right)}}.$$

Under the null hypothesis, h_a^2 was set to be 0, and β_m became 0. Liabilities for each individual were generated from the sum of the main genetic effects, polygenic effects, common environmental effects, and random errors, and they were transformed to being affected if they were larger than the threshold; otherwise, they were considered to be unaffected. The threshold was set to generate the assumed prevalence q. Disease prevalences are sometimes different between males and females, and this was considered by setting different prevalence rates for males and females in our simulations. Randomly selected families can have very few affected individuals, which leads to the large false negative finding. Therefore, we considered some ascertainment strategies. That is, families with less than two affected grandchildren were excluded from the simulation studies, and sampling was repeated until the desired number of families was obtained.

We also evaluated the proposed methods in the presence of population substructure. We assumed two underlying sub-populations, and each founder was randomly assigned to one of two sub-populations. The polygenic effect, common environmental effect, and random errors were generated with the same model used in the absence of population substructure. However, the phenotypic means of liabilities between two sub-populations were varied by 0.5. The allele frequencies for the two subpopulations were generated with the Balding-Nichols model [Balding and Nichols 1995]. We first generate p_m for global population MAF from $U(0, 0.05)$. Then, if we let F_{ST} denote Wright's F_{ST}, MAFs for two sub-populations were independently sampled from $Beta(p_m(1 - F_{ST})/F_{ST}, (1 - p_m)(1 - F_{ST})/F_{ST})$. F_{ST} was assumed to 0, 0.005, 0.01, and 0.05.

Results

Evaluation with simulated data

We estimated type-1 error rates and powers of the proposed methods, and results from the proposed method were compared with PedGene-Burden and PedGene-Kernel statistics [Schaid, et al. 2013]. In particular, PedGene-Burden and PedGene-Kernel cannot handle S-XCI model and they were not considered for S-XCI model. We considered five different extended family structures (A-1) – (A-5) as shown in Figure 1. We assumed that there were 200 extended families and 30 rare variants in each gene. Empirical type-1 errors were calculated at the 0.05 and 0.01 significance levels with 5,000 replicates for dichotomous phenotypes. Supplementary Tables S1 and S2 show that type-1 error estimates of our proposed methods consistently preserved the nominal significance levels for any biological expression process, whereas the statistical validity of PedGene-Burden and PedGene-Kernel depends on family structure and type-1 error estimates of PedGene-Burden are violated for (A-1), (A-2), (A-4), and (A-5) of E-XCI. Supplementary Tables S3 – S6 show the type-1 error estimates when disease prevalences for males and females are different. Disease prevalences were set to be 0.36 and 0.12 for males and females respectively in Supplementary Tables 3-4, and 0.12 and 0.36 in Supplementary Tables 5-6. Results show that the proposed methods always preserve the nominal significance levels. However type-1 error estimates of PedGene-Burden and PedGene-Kernel for E-XCI model setting consistently preserved the nominal significance levels.

In order to evaluate statistical efficiency, we considered five different extended family structures (A-1) – (A-5), and calculated the empirical power estimates for each. We assumed that there are 30 rare variants in each gene and 20 of them are causal. The number of deleterious causal rare variants was assumed to be 10, 12, 16, or 20. We assumed that h_a^2 was 0.01 and empirical power values at the 0.05 significance level were estimated with 5,000 replicates. Figure 2, Supplementary Fig. S2 and S3 show that ***FARVAT-XB*** was the most powerful statistic if all risk variants are deleterious. If half of rare causal variants were deleterious and the other rare causal variants were protective,***FARVAT-XC*** was the most powerful statistic. ***FARVAT-XO*** and ***FARVAT-XD*** were not always most efficient, but differences of power estimates among ***FARVAT-XO***, ***FARVAT-XD*** and the most efficient statistic were always small. It should be noted that ***FARVAT-XD*** is robust against the choice of mis-specified *d*. Supplementary Fig. S1 shows that PedGene-Burden is the most efficient statistic under E-XCI if all rare causal variants were deleterious, but it should be noted that empirical type-1 errors from PedGene-Burden were violated.

Evaluation with simulated data in the presence of population substructure

We estimated the type-1 error rate and power for the proposed methods in the presence of population substructure, and compared them to the same statistics from PedGene-Burden and PedGene-Kernel. In our proposed method, the presence of population substructure can be handled by adjusting the phenotypes with an EIGENSTRAT-based approach [Schaid, et al. 2013; Won, et al. 2012]. Specifically, principal component (PC) scores were estimated from the genetic relation matrix [Price, et al. 2006], and phenotypes were regressed on PC scores with the linear mixed model, which considers the correlation between family members. Residuals were then utilized as t_{ij} for the proposed methods. The type-1 error estimates for trios were calculated at the 0.05 and 0.01 significance levels with 5,000 replicates. We assumed that there were 30 rare variants available in a gene and family structure (B-1) and (B-2) in Figure 1. Supplementary Table S7 shows inflation of type-1 error estimates for all methods unless phenotypes are adjusted with PC scores, and, in particular, PedGene-Kernel has the largest bias of type-1 error estimates.

The statistical efficiency was also evaluated with 5,000 replicates at the 0.05 significance level in the presence of population substructure. We assumed that h_a^2 is 0.05, and that there are 30 rare variants in a gene. Twenty rare variants were assumed to be causal, and each causal variant can have either deleterious or protective effects on phenotypes. Figure 3 shows that ***FARVAT-XB*** was the most efficient when all rare causal variants are deleterious, and PedGene-Kernel was the most powerful if 50% of rare causal variants was deleterious. ***FARVAT-XO*** and ***FARVAT-XD*** are not always the most efficient, but their power loss when compared to the most efficient statistic is always small.

Evaluation of robustness against biological expression process

The gene expression process of X-linked variants is usually unknown, and the misspecified gene expression process may affect the performance of the proposed methods. We evaluated the robustness of the proposed methods with simulated data for (A-3) family structure. The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01

significance levels and supplementary Table S8 shows that type-1 error estimates of ***FARVAT-XO*** and ***FARVAT-XD*** consistently preserved the nominal significance levels. For evaluation of statistical powers, h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Supplementary Fig. S5 shows that ***FARVAT-XO*** with correctly specified biological model is the most efficient, but if it is misspecified, the power loss is usually substantial. ***FARVAT-XD*** is not the most efficient but the difference of its statistical powers with those for ***FARVAT-XO*** with correctly specified biological model is very small. Therefore, we can conclude that the performance of ***FARVAT-XO*** is affected by choice of *d*, and ***FARVAT-XD*** is generally a robust choice for various biological processes.

Application to COPD data

The proposed methods were applied to rare variant association analyses of COPD using families from the Boston Early-Onset COPD Study with whole exome sequencing. Using moderate COPD or greater (FEV1 < 80% predicted with FEV1/FVC < 0.7) to define affection status, there were 64 unaffected males, 83 unaffected females, 55 affected males, and 100 affected females. There were 49 families and each family had at least two affected individuals. The whole exome of all individuals was sequenced with a Nimblegen V2 capture and Illumina platform. Sequencing data were preprocessed with the Genome Analysis ToolKit [McKenna, et al. 2010]. SNVs with Mendelian transmission errors, missing call rates (>1%), significant deviation from Hardy–Weinberg equilibrium ($P<10^{-8}$), read depth less than the average (12), and minor allele count of all variants in each gene (<5) were excluded. Seven genes in pseudo-autosomal regions and 186 genes with a single rare variant were excluded from our analyses. In total, we analyzed 629 rare variants in 183 genes on the X chromosome. There were 35,326 common autosomal variants with a MAF larger than 0.05, and they were utilized to calculate the genetic relationship matrix. Supplementary Figure S4 shows the genetic relationships of the dataset on the first five PC scores. Phenotypes were regressed with age, pack years, height, and 5 PC scores from the EIGENSTRAT method [Price, et al. 2006], and residuals were utilized as response variables to provide robustness of the proposed methods against population substructure. Figures 4 and 5 show quantile-quantile (QQ) plots of PedGene-Burden, PedGene-Kernel, and the proposed methods. QQ plots for PedGene-Burden and PedGene-Kernel show some evidence about inflation under random XCI and E-XCI, whereas the proposed methods are consistently valid. The most significant results were summarized in Table 2. The 0.05 exome-wide significant level adjusted by Bonferroni correction is 2.7E-04, and *q*-values [Storey 2002] were also provided in Table 2. Table 2 showed one exome-wide significant gene, *CXorf59* gene, with PedGene-Kernel for random XCI. However some inflation of results from PedGene-Kernel was confirmed with QQ plots and is not clear whether this significant association is valid. Some other promising results are also summarized in Table 2 and the second most significant results were obtained for the synovial sarcoma on X chromosome 5 (*SSX5*) gene using the proposed method. The significant association of *SYT-SSX* fusion gene with primary synovial sarcoma of the lung was reported [Hisaoka, et al. 1999], and the expression of SSX family genes (*SSX1*, *SSX2*, *SSX4*, and *SSX5*) were known to be related with lung cancer [Tureci, et al. 1998]. Furthermore, the *COL4A6*

isoform have been shown to be more highly expressed in lung [Hudson, et al. 1993] and these significant results will be investigated as further studies.

Discussion

X-linked genes contribute to various biological mechanisms, including sexual dimorphism [Carrel and Willard 2005; Ober, et al. 2008; Tarpey, et al. 2009]. However, the complex biological processes associated with the expression of X-linked genes, such as random XCI, E-XCI, and S-XCI, complicate genetic association analyses with X-linked genes. Several methods for rare X-linked variants association analyses have been developed, but most cannot account for biologically plausible models. The limited discovery of significantly associated X-linked variants may be partially attributable to the absence of statistically efficient methods for detecting X-linked variants, and efficient analytical strategies for X-linked variants have been proposed as a potential mechanism to alleviate so-called “missing heritability” problems [Maher 2008; Manolio, et al. 2009].

In this report, we proposed a novel method for family-based association test of X-linked genes (***FARVATX***), which can accommodate random XCI, E-XCI, and S-XCI. The performance of ***FARVATX*** was evaluated with simulated data. We assumed that the magnitude of X-linked gene expression differed by gender and that the proportion of males and females in each family was different. The results from the simulation studies showed that PedGene-Burden and PedGene-Kernel statistics suffer from inflation of the type 1 error rate if the proportions of males and females are different or population substructure is present. However, ***FARVATX*** preserves the nominal significance level in both the absence and presence of population substructure.

Furthermore, ***FARVATX*** is computationally less intensive than other available methods. Its application to sequencing data for COPD was completed within an hour. ***FARVATX*** software supports various input file formats, including plink and variant call format files, and multi-threaded analyses can be automatically conducted. The software for the proposed methods is written in C++ and can be downloaded from http://healthstat.snu.ac.kr/software/farvatx/.

Despite the analytical flexibility of the proposed methods, there are still some limitations. First, we found that the proposed methods are slightly conservative unless the sample size is sufficiently large, and it has been shown that small sample size adjustments by using resampling method leads to additional power improvement [Lee, et al. 2012a]. Second, the statistical power depends on the definition of rare variants, but it is still unclear. A variable threshold approach [Price, et al. 2010] that exhaustively searches the optimal MAF threshold may be a useful option for addressing this issue, and further extensions for the proposed methods are necessary. Third, the proposed methods assume that MAFs are same for males and females under the null hypothesis, and effects of each genetic variant for males and females are similar under the alternative hypothesis. If these are not satisfied, the false negative finding rates for the former and false positive findings rates for the latter cannot be controlled, and males and females should be separately analyzed. These problems will be investigated in future studies.

The recent rapid improvement of sequencing technology provides the opportunity to identify rare X-linked variants associated with complex human diseases. However, our understanding of sex-specific genetic architecture and the biological processes associated with the expression of X-linked genes is still limited, and statistical methodology development to uncover them is necessary. The proposed methods may help us identify additional rare X-linked variants associated with complex traits, thereby leading to about a better understanding of the underlying biological processes associated with X-linked genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant (2012R1A3A2026438) and by the Bio & Medical Technology Development Program of the NRF grant (2013M3A9C4078158) to TP, and by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559), the Industrial Core Technology Development Program (10040176, Development of Various Bioinformatics Software using Next Generation Bio-data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea), Basic Science Re-search Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2013R1A1A2010437) to SW, NIH P01 HL105339, R01 HL075478, P01 HL114501 (EKS), and R01 HL113264 (EKS and MHC). Sequencing for the Boston Early-Onset COPD Study was provided by the University of Washington Center for Mendelian Genomics (UW CMG) and was funded by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute grant 1U54HG006493 to Drs. Debbie Nickerson, Jay Shendure, and Michael Bamshad.

References

Abkowitz JL, Taboada M, Shelton GH, Catlin SN, Guttorp P, Kiklevich JV. An X chromosome gene regulates hematopoietic stem cell kinetics. Proc Natl Acad Sci U S A. 1998; 95(7):3862–6. [PubMed: 9520458]

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S. Medical sequencing at the extremes of human body mass. Am J Hum Genet. 2007; 80(4):779–91. others. [PubMed: 17357083]

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(3):493–9. [PubMed: 16909387]

Armitage P. Tests for Linear Trends in Proportions and Frequencies. Biometrics. 1955; 11(3):375–386.

Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N, Carlson C, de Denus S, Dube MP, Haessler J. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet. 2014; 46(6):629–34. others. [PubMed: 24777453]

Balding DJ, Nichols RA. A Method for Quantifying Differentiation between Populations at Multi-Allelic Loci and Its Implications for Investigating Identity and Paternity. Genetica. 1995; 96(1-2):3–12. [PubMed: 7607457]

Belmont JW. Genetic control of X inactivation and processes leading to X-inactivation skewing. Am J Hum Genet. 1996; 58(6):1101–8. [PubMed: 8651285]

Blakemore AI, Meyre D, Delplanque J, Vatin V, Lecoeur C, Marre M, Tichet J, Balkau B, Froguel P, Walley AJ. A rare variant in the visfatin gene (NAMPT/PBEF1) is associated with protection from obesity. Obesity (Silver Spring). 2009; 17(8):1549–53. [PubMed: 19300429]

Brown CJ, Greally JM. A stain upon the silence: genes escaping X inactivation. Trends Genet. 2003; 19(8):432–8. [PubMed: 12902161]

Brown MB. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. Biometrics. 1975; 31(4):987–992.

Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y, Maragh M, Gilliland DG. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood. 1996; 88(1):59–65. [PubMed: 8704202]

Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005; 434(7031):400–4. [PubMed: 15772666]

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(10):1209–14. [PubMed: 16219543]

Choi S, Lee S, Cichon S, Nothen MM, Lange C, Park T, Won S. FARVAT: a family-based rare variant association test. Bioinformatics. 2014; 30(22):3197–205. [PubMed: 25075118]

Clayton D. Testing for association on the X chromosome. Biostatistics. 2008; 9(4):593–600. [PubMed: 18441336]

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305(5685):869–72. [PubMed: 15297675]

Gaukrodger N, Mayosi BM, Imrie H, Avery P, Baker M, Connell JM, Watkins H, Farrall M, Keavney B. A rare variant of the leptin gene has large effects on blood pressure and carotid intima-medial thickness: a study of 1428 individuals in 248 families. J Med Genet. 2005; 42(6):474–8. [PubMed: 15937081]

Hisaoka M, Hashimoto H, Iwamasa T, Ishikawa K, Aoki T. Primary synovial sarcoma of the lung: report of two cases confirmed by molecular detection of SYT-SSX fusion gene transcripts. Histopathology. 1999; 34(3):205–10. [PubMed: 10217560]

Hudson BG, Reeders ST, Tryggvason K. Type IV collagen: structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis. J Biol Chem. 1993; 268(35):26033–6. [PubMed: 8253711]

Knudsen GP, Pedersen J, Klingenberg O, Lygren I, Orstavik KH. Increased skewing of X chromosome inactivation with age in both blood and buccal cells. Cytogenet Genome Res. 2007; 116(1-2):24–8. [PubMed: 17268174]

Lange C, Laird NM. Power calculations for a general class of family-based association tests: dichotomous traits. Am J Hum Genet. 2002; 71(3):575–84. [PubMed: 12181775]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team NGESP-ELP. Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012a; 91(2):224–37. [PubMed: 22863193]

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012b; 13(4):762–75. [PubMed: 22699862]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. [PubMed: 18691683]

Lyon MF. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature. 1961; 190:372–3. [PubMed: 13764598]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. [PubMed: 19214210]

Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456(7218):18–21. [PubMed: 18987709]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. others. [PubMed: 19812666]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20(9):1297–1303. others. [PubMed: 20644199]

McPeek MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004; 60(2):359–67. [PubMed: 15180661]

Minks J, Robinson WP, Brown CJ. A skewed view of X chromosome inactivation. J Clin Invest. 2008; 118(1):20–3. [PubMed: 18097476]

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7(3):e1001322. [PubMed: 21408211]

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324(5925):387–9. [PubMed: 19264985]

Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. Nat Rev Genet. 2008; 9(12):911–22. [PubMed: 19002143]

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(1):168–73. [PubMed: 12068376]

Price AL, Kryukov GV, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. [PubMed: 20471002]

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(8): 904–9. [PubMed: 16862161]

Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997; 53(4):1253–61. [PubMed: 9423247]

Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol. 2013; 37(5):409–18. [PubMed: 23650101]

Shapiro LJ, Mohandas T, Weiss R, Romeo G. Non-inactivation of an x-chromosome locus in man. Science. 1979; 204(4398):1224–6. [PubMed: 156396]

Sharp A, Robinson D, Jacobs P. Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet. 2000; 107(4):343–9. [PubMed: 11129333]

Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology. 2002; 64:479–498.

Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet. 2009; 41(5):535–43. others. [PubMed: 19377476]

Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007; 81(2):321–37. [PubMed: 17668381]

Thornton T, Zhang Q, Cai X, Ober C, McPeek MS. XM: association testing on the X-chromosome in case-control samples with related individuals. Genet Epidemiol. 2012; 36(5):438–50. [PubMed: 22552845]

Tureci O, Chen YT, Sahin U, Gure AO, Zwick C, Villena C, Tsang S, Seitz G, Old LJ, Pfreundschuh M. Expression of SSX genes in human tumors. Int J Cancer. 1998; 77(1):19–23. [PubMed: 9639388]

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(6):483–93. [PubMed: 25043884]

Won S, Lange C. A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. Stat Med. 2013; 32(25):4482–98. [PubMed: 23740776]

Won S, Lu Q, Bertram L, Tanzi RE, Lange C. On the meta-analysis of genome-wide association studies: a robust and efficient approach to combine population and family-based studies. Hum Hered. 2012; 73(1):35–46. [PubMed: 22261799]

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(3):e17873. [PubMed: 21445353]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. [PubMed: 21737059]

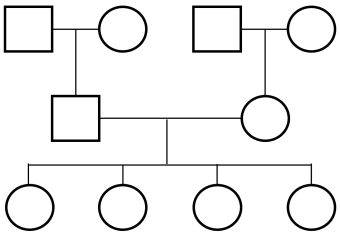
Zheng G, Freidlin B, Li ZH, Gastwirth JL. Choice of scores in trend tests for case-control studies of candidate-gene associations. Biometrical Journal. 2003; 45(3):335–348.

Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. Genet Epidemiol. 2007; 31(8):834–43. [PubMed: 17549761]

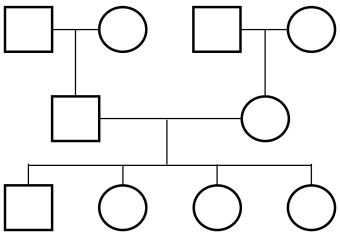
Zhu Y, Xiong M. Family-Based Association Studies for Next-Generation Sequencing. American Journal of Human Genetics. 2012; 90(6):1028–1045. [PubMed: 22682329]

(A) Extended families

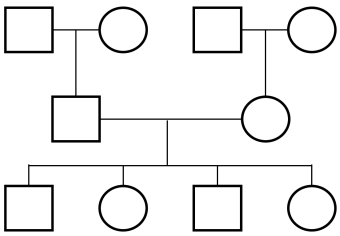
(A-1)



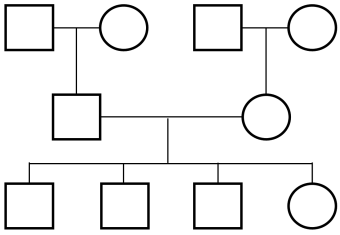
(A-2)



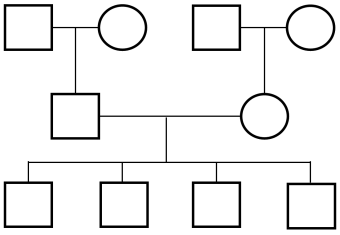
(A-3)



(A-4)

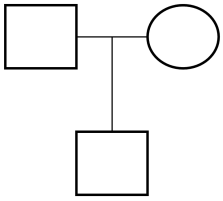


(A-5)



(B) Trios

(B-1)



(B-2)

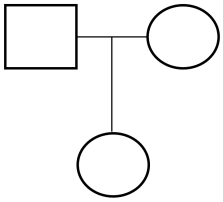


Figure 1.
Family structures considered in our simulation studies.

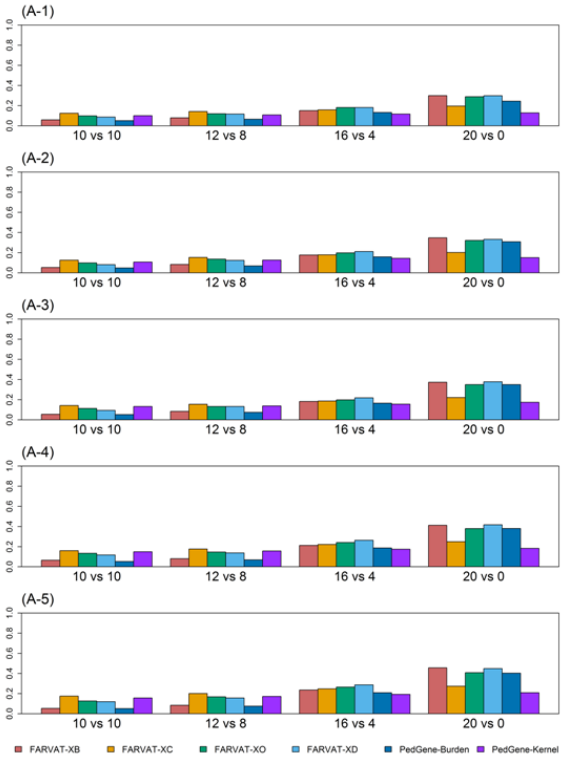


Figure 2. Empirical power estimates for random XCI

Empirical powers were calculated for five different extended family structures (A-1) – (A-5). h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.

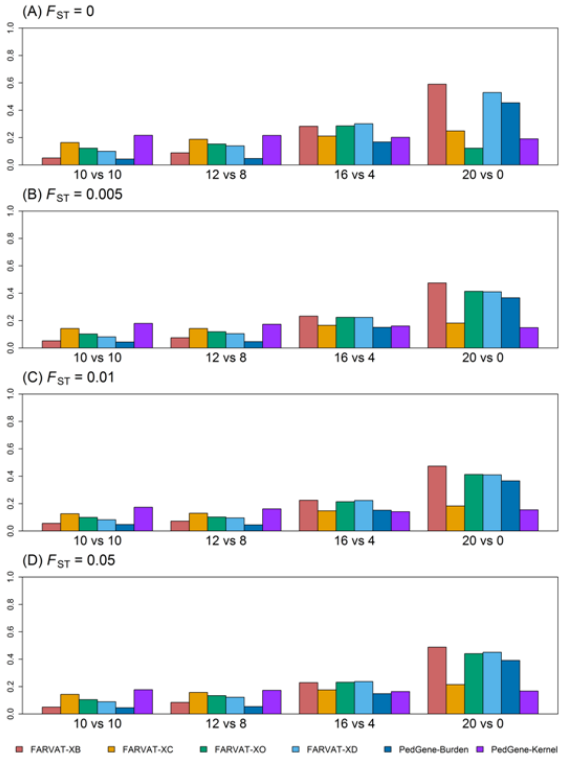


Figure 3. Empirical power estimates for random XCI in the presence of population substructure

Empirical powers were calculated for two different trio structures (B-1) and (B-2). h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.

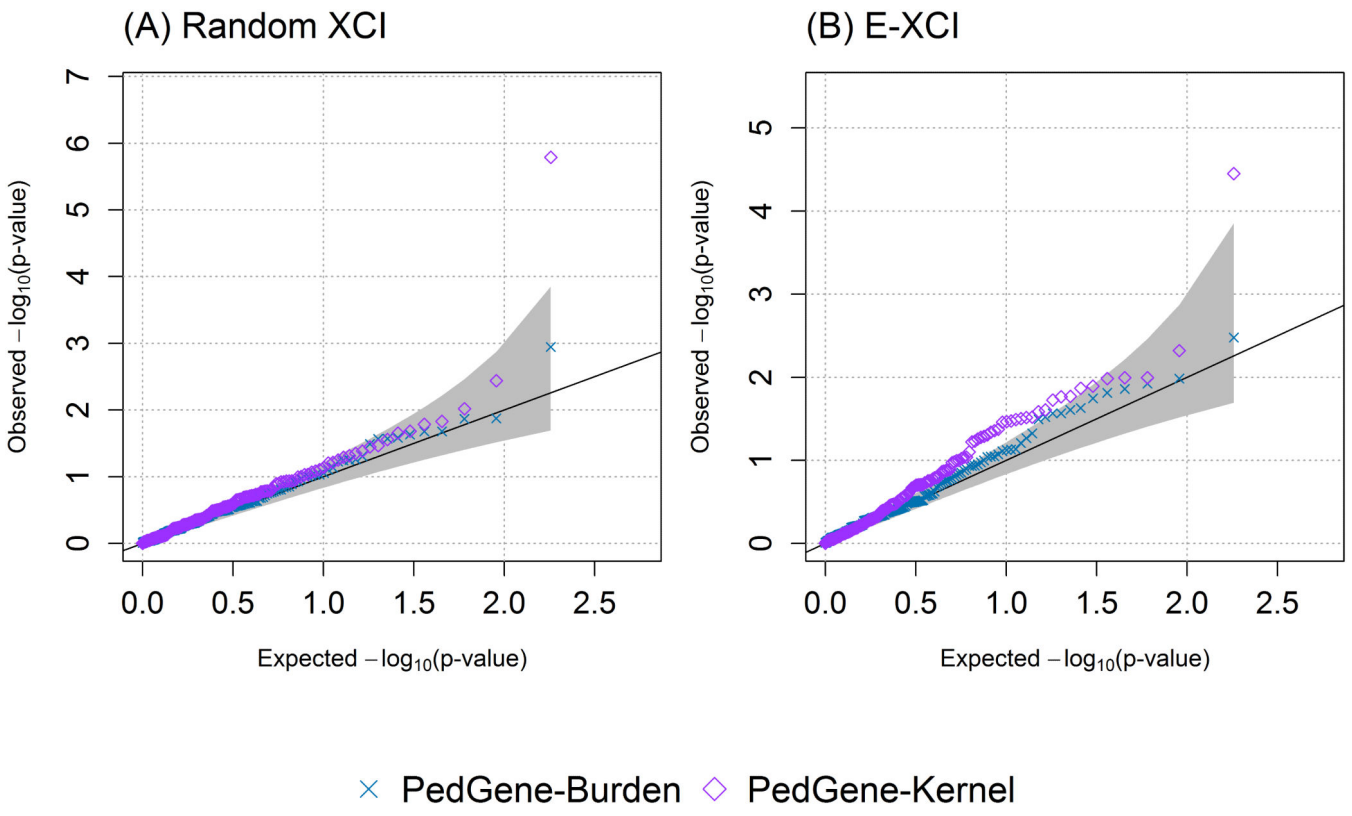


Figure 4. QQ-plots of results from rare variant association analyses of COPD

QQ-plots are provided for PedGene-Burden, and PedGene-Kernel, and their 95% confidence interval is provided. Age, Pack-years of smoking, height, and 5 PCs were included as covariates for the linear mixed model and BLUP was utilized as offset.

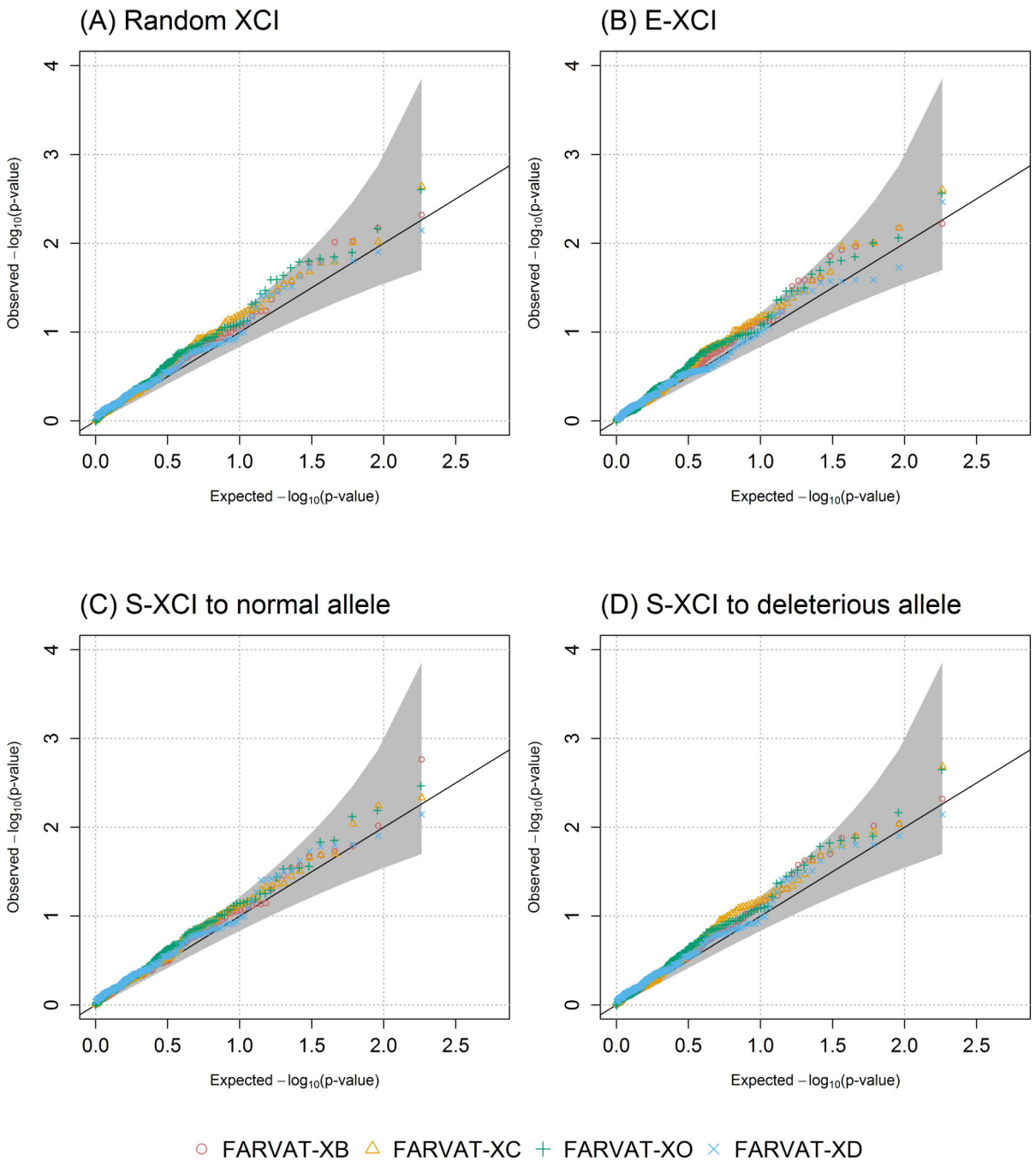


Figure 5. QQ plots of results from rare variant association analyses of COPD

QQ plots are provided for ***FARVAT-XB***, ***FARVAT-XC, FARVAT-XO***, and ***FARVAT-XD***, and their 95% confidence interval is provided. Age, Pack-years of smoking, height, and 5 PCs were included as covariates for the linear mixed model, and BLUP was utilized as offset.

Table 1

X chromosomal and autosomal kinship coefficients for two individuals in a nuclear family.

Relationship of individuals ij and ij'	$\pi_{ij,ij'}$ (X chromosome)	$\pi_{ij,ij'}$ (autosome)
Brother & Brother	1/4	1/4
Sister & Sister	3/4	1/4
Brother & Sister	1/4	1/4
Mother & Son	1/2	1/4
Mother & Daughter	1/2	1/4
Father & Son	0	1/4
Father & Daughter	1/2	1/4

Table 2

Most significant results from rare variant association analyses of COPD data.

Models	GENE	*M*	MAC		*FARVAT -XB*		*FARVAT -XC*		*FARVAT -XO*		*FARVAT -XD*		PedGene -Burden		PedGene -Kernel	
			Aff	Un	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*
XCI	*CXorf59*	2	2	7	0.165	0.798	0.016	0.590	0.026	0.425	0.039	0.552	0.001	0.205	1.6 E-06	2.9 E-04
	MTMR8	6	5	1	0.005	0.442	0.063	0.607	0.007	0.420	0.013	0.552	0.180	0.841	0.191	0.734
	SSX5	2	5	2	0.512	0.895	0.002	0.419	0.003	0.420	0.035	0.552	0.026	0.547	0.093	0.698
E-XCI	*CXorf59*	2	2	7	0.372	0.891	0.120	0.662	0.174	0.670	0.003	0.624	0.015	0.484	3.6 E-05	6.4 E-03
	ELF4	4	13	5	0.919	0.957	0.934	0.955	0.964	0.969	0.784	0.903	0.003	0.484	0.034	0.330
	MTMR8	6	5	1	0.006	0.423	0.051	0.632	0.009	0.493	0.377	0.865	0.077	0.715	0.030	0.330
	SSX5	2	5	2	0.761	0.957	0.003	0.397	0.003	0.493	0.026	0.637	0.025	0.484	0.059	0.387
S-XCI to normal allele	*COL4A6*	7	18	21	0.02	0.316	0.022	0.677	0.003	0.459	0.007	0.552				
	CXorf59	2	2	7	0.094	0.783	0.005	0.524	0.008	0.459	0.039	0.552				
S-XCI to deleterious allele	*MTMR8*	6	5	1	0.005	0.482	0.050	0.579	0.007	0.424	0.013	0.552				
	SSX5	2	5	2	0.650	0.917	0.002	0.378	0.002	0.407	0.035	0.552				

Notes. The significant results for ***FARVAT-XB***, ***FARVAT-XC***, ***FARVAT-XO***, ***FARVAT-XD***, PedGene-Burden, and PedGene-Kernel are provided. The 0.05 exome-wide significant level adjusted by Bonferroni correction is 2.7E-04, and *q*-values [Storey 2002] are provided. *M* indicates the number of rare variant in a gene, and MAC indicates the minor allele counts.