Xu *et al. BMC Genetics* (2018) 19:109
https://doi.org/10.1186/s12863-018-0694-8

BMC Genetics

METHODOLOGY ARTICLE Open Access

A statistical measure for the skewness of X chromosome inactivation based on family trios

Si-Qi Xu[1], Yu Zhang[1], Peng Wang[1], Wei Liu[1], Xian-Bo Wu[2*] and Ji-Yuan Zhou[1*]

Abstract

Background: X chromosome inactivation (XCI) is an important gene regulation mechanism in females to equalize the expression levels of X chromosome between two sexes. Generally, one of two X chromosomes in females is randomly chosen to be inactivated. Nonrandom XCI (XCI skewing) is also observed in females, which has been reported to play an important role in many X-linked diseases. However, there is no statistical measure available for the degree of the XCI skewing based on family data in population genetics.

Results: In this article, we propose a statistical approach to measure the degree of the XCI skewing based on family trios, which is represented by a ratio of two genotypic relative risks in females. The point estimate of the ratio is obtained from the maximum likelihood estimates of two genotypic relative risks. When parental genotypes are missing in some family trios, the expectation-conditional-maximization algorithm is adopted to obtain the corresponding maximum likelihood estimates. Further, the confidence interval of the ratio is derived based on the likelihood ratio test. Simulation results show that the likelihood-based confidence interval has an accurate coverage probability under the situations considered. Also, we apply our proposed method to the rheumatoid arthritis data from USA for its practical use, and find out that a locus, rs2238907, may undergo the XCI skewing against the at-risk allele. But this needs to be further confirmed by molecular genetics.

Conclusions: The proposed statistical measure for the skewness of XCI is applicable to complete family trio data or family trio data with some paternal genotypes missing. The likelihood-based confidence interval has an accurate coverage probability under the situations considered. Therefore, our proposed statistical measure is generally recommended in practice for discovering the potential loci which undergo the XCI skewing.

Keywords: X chromosome inactivation, Skewing, Family trio, Ratio estimate

Background

Many human diseases are associated with genes on X chromosome, such as asthma, autoimmune diseases, cancers, some neurological and psychiatric diseases [1–5]. Most of these X-linked diseases often exhibit sex-specific patterns of susceptibility due to the difference in the number of copies of X chromosome between two sexes. Females have two copies of X chromosome whereas there is only one copy in males. To equalize the expression levels of X chromosome between sexes, dosage compensation is achieved by an important gene regulation mechanism in mammalian females, X chromosome inactivation (XCI), which results in expression silencing of one of two X chromosomes in females [6]. Up to 75% genes on X chromosome are subject to XCI, while there are about 15% escaping from inactivation and expressed from both X chromosomes, and the remaining 10% show variable inactivation patterns in different human cell lines [7].

During the process of XCI, one of two X chromosomes in females is chosen to be inactivated in a random way. This means that roughly 50% of cells in females have the paternal X chromosome expressed, while the others express the maternal one. Although random XCI occurs

* Correspondence: wuxb1010@hotmail.com; zhoujiyuan5460@hotmail.com
[2]Guangdong Provincial Key Laboratory of Tropical Disease Research, Department of Epidemiology, School of Public Health, Southern Medical University, Guangzhou, China
[1]State Key Laboratory of Organ Failure Research, Ministry of Education and Guangdong Provincial Key Laboratory of Tropical Disease Research, Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

commonly, the XCI skewing also takes place in females, which is defined as the phenomenon of nonrandom inactivation that one of X chromosomes is selected to be silenced with a probability deviating from 50% [8]. Generally, the skewness of XCI is caused by a second selection mechanism. When the mutation on X chromosome affects the survival and proliferation of cells, the amount of cells carrying an active mutant X chromosome will become larger or smaller than that of cells with an active wild-type X chromosome, which thus leads to the skewness of XCI [9, 10]. Negative selection, where the mutation gives a growth disadvantage to cells, frequently happens in female carriers with X-linked diseases, such as mental retardation disorders, Wiskott-Aldrich syndrome and X-linked severe combined immunodeficiency [11–13]. On the other hand, when the mutation provides a growth advantage to cells, positive selection occurs and can result in some diseases, such as adrenoleukodystrophy and breast cancer [14, 15].

In genetic association studies on X chromosome, Clayton [16] first took XCI into consideration. Due to XCI, the genotypic effect of homozygous females can be treated the same as that of hemizygous males. Therefore, the genotypic scores are given to be 0, 1 and 2 corresponding to three genotypes at a diallelic locus on X chromosome in females, and 0 and 2 corresponding to two genotypes in males. However, Wang et al. [17] pointed out that such coding strategy only considers the situation of random XCI, but ignores the skewness of XCI and escape from XCI. To account for all possible situations of XCI, Wang et al. suggested that the genotypic score for the heterozygous females, denoted by γ, can be any possible values between 0 and 2. Under XCI, the value of γ reflects the degree of inactivation skewing, with $\gamma/2$ representing the proportion of cells having the mutant allele active. As such, $\gamma = 1$ stands for random XCI, while γ between 1 and 2 indicates the XCI skewing toward the mutant allele and γ between 0 and 1 denotes the XCI skewing against the mutant allele. For example, $\gamma = 0.5$ means that the skewness of XCI is against the mutant allele with 25% cells expressing the mutant allele and the other 75% cells expressing the normal allele. On the other hand, in molecular genetics, the XCI skewing pattern can be identified by assays taking advantage of differential methylation between the active and inactive X chromosomes or mRNA transcription in cells [18–20]. However, since the XCI pattern always varies among cell lines [21, 22], these assays, which usually use cells from only a few tissues to investigate the XCI patterns, cannot present the status of the whole body [10]. Further, there is no statistical measure available for detecting the XCI skewing pattern in population genetics as yet.

Therefore, in this article, we give the expression of γ for family trios with both parents and one affected daughter in the presence of association between the disease and genotypes. In fact, γ is a function of two genotypic relative risks (GRRs) in females. In addition, we obtain the point estimate of γ from the maximum likelihood estimates (MLEs) of the GRRs. When parental genotypes are missing in some family trios, the expectation-conditional-maximization (ECM) algorithm [23] is used to obtain the corresponding MLEs. Further, the confidence interval (CI) of γ is derived based on a likelihood ratio test (LRT). Finally, simulation study is conducted to investigate the performance of our proposed method. The simulation results show that the proposed likelihood-based CI has an accurate coverage probability under the situations considered. For practical use, we apply our proposed method to the rheumatoid arthritis (RA) data from USA.

Methods

Notations

Consider an X-linked diallelic locus with mutant allele A and normal allele a. Let p_m and $q_m = 1 - p_m$ denote the allele frequencies of A and a in males, respectively. Suppose that p_f is the allele frequency of A and ρ is the inbreeding coefficient in females. Then, the frequencies of genotypes aa, Aa and AA in females are respectively $g_0 = (1 - p_f)^2 + \rho p_f(1 - p_f)$, $g_1 = 2(1 - \rho)p_f(1 - p_f)$ and $g_2 = p_f^2 + \rho p_f(1 - p_f)$. Note that Hardy-Weinberg equilibrium holds in the population under study when $\rho = 0$ and $p_m = p_f$. Let f_0, f_1 and f_2 respectively represent the penetrances in females with genotypes aa, Aa and AA. The GRRs in females are defined as $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$.

Relationship between penetrances and XCI skewing in females

Let the genotypic scores be 0, γ and 2 corresponding to females with genotypes aa, Aa and AA, respectively, where $\gamma \in [0, 2]$ represents the XCI skewing pattern. We assume that a generalized genetic model holds [24, 25], which is defined as $f_0 \leq f_1 \leq f_2$ (i.e., $1 \leq \lambda_1 \leq \lambda_2$) with at least one inequality being strict, in the presence of association between the disease and genotypes in females. If f_1 is unknown, then it can be expressed as a function of γ, denoted by $f_1(\gamma)$. To derive the expression of $f_1(\gamma)$, let $f_1'(\gamma)$ be the first order derivative of $f_1(\gamma)$ with respect to γ. As such, $f_1(\gamma)$ can be approximated by a first order Taylor expansion around $\gamma = 1$ as follows,

$$f_1(\gamma) \approx f_1(1) + f_1'(1)(\gamma - 1). \tag{1}$$

On the other hand, when the XCI skewing is completely against A, we have γ=0 and $f_1 = f_0$. So, from Eq. (1), $f_0 = f_1(0) \approx f_1(1) - f_1'(1)$. Similarly, when the XCI skewing is completely toward A, we have γ=2 and $f_1 = f_2$.

Then $f_2 = f_1(2) \approx f_1(1) + f_1'(1)$. Hence, $f_1(1) = (f_2 + f_0)/2$ and $f_1'(1) = (f_2 - f_0)/2$. Therefore, Eq. (1) turns to be

$$f_1(\gamma) \approx f_0 + \frac{\gamma(f_2 - f_0)}{2}. \tag{2}$$

From Eq. (2), we notice that f_1 is around the midpoint between f_0 and f_2 under random XCI (i.e., γ=1). Actually, Eq. (2) means that the penetrance for heterozygous females is approximately linear in the genotypic score γ around γ=1.

Further, if f_1 is known, then we can obtain γ from Eq. (2) as follows, which is a function of f_0, f_1 and f_2, or λ_1 and λ_2,

$$\gamma \approx \frac{2(f_1 - f_0)}{f_2 - f_0} = \frac{2(\lambda_1 - 1)}{\lambda_1 - 1}.$$

Note that the value of γ attains its maximum ($\gamma = 2$) when $\lambda_1 = \lambda_2 \neq 1$, and $\gamma = 0$ when $\lambda_1 = 1$ and $\lambda_2 \neq 1$. Assume that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the MLEs of λ_1 and λ_2, respectively. Then, the point estimate of γ, $\hat{\gamma}$, can be obtained by $2(\hat{\lambda}_1 - 1)/(\hat{\lambda}_2 - 1)$.

MLEs of λ_1 and λ_2 using family trios without missing genotypes

Here we only include the trios with affected daughters in the analysis. Male offspring are not investigated because they are not informative of λ_1 and λ_2. Firstly, we consider complete family trios, each with both typed parents and an affected typed daughter. Let *F*, *M* and *C* represent the numbers of allele *A* in father, mother and daughter, respectively, and *D* denote that the daughter is affected. Eight possible types of *FMC* (i.e., 000, 010, 011, 021, 101, 111, 112 and 122) together with the corresponding probabilities $P(FM)$, $P(C|\, FM)$ and $P(FMC|\, D)$ are shown in Table 1. Let Ω be the set of the eight possible types of *FMC* listed in Table 1 and then $P(FMC|\, D)$ is derived as

$$P(FMC|D) = \frac{P(FM)P(C|FM)P(D|C)}{\sum_{F'M'C'\in\Omega} P(F'M')P(C'|F'M')P(D|C')}. \tag{3}$$

Equation (3) holds when the disease status of a daughter is only related to her own genotype, and $P(C|\, FM)$ is determined by Mendelian transmission, which is equal to 0.5 for heterozygous mother and 1 otherwise. Assume that $P(FM) = P(F)P(M)$, and divide the numerator and denominator of Eq. (3) by f_0. Then, $P(FMC|\, D)$ for each trio type can be written as the last column in Table 1, where $R = q_m g_0 + 0.5 q_m g_1 (1 + \lambda_1) + q_m g_2 \lambda_1 + p_m g_0 \lambda_1 + 0.5 p_m g_1 (\lambda_1 + \lambda_2) + p_m g_2 \lambda_2$. The detailed derivation of $P(FMC|\, D)$ in Table 1 is given in Additional file 1: Appendix A.

Table 1 Eight types of possible family trios and the corresponding probabilities

FMC	*P(FM)*	*P(C| FM)*	*P(FMC| D)*
000	$q_m g_0$	1	$q_m g_0/R$
010	$q_m g_1$	0.5	$0.5 q_m g_1/R$
011	$q_m g_1$	0.5	$0.5 q_m g_1 \lambda_1/R$
021	$q_m g_2$	1	$q_m g_2 \lambda_1/R$
101	$p_m g_0$	1	$p_m g_0 \lambda_1/R$
111	$p_m g_1$	0.5	$0.5 p_m g_1 \lambda_1/R$
112	$p_m g_1$	0.5	$0.5 p_m g_1 \lambda_2/R$
122	$p_m g_2$	1	$p_m g_2 \lambda_2/R$

Since we find that it is more convenient to directly estimate g_0 and g_1 rather than ρ and p_f, we let $\theta = (p_m, g_0, g_1, \lambda_1, \lambda_2)^T$ be the parameter vector of interest. As such, the log-likelihood function of the observed data conditional on the daughter being affected is given by

$$\ln L(\theta) = \sum_{FMC\in\Omega} n_{FMC} \ln P(FMC|D),$$

where n_{FMC} is the number of the family trios of type *FMC*. To obtain the MLE of θ, numerical methods, such as Newton-Raphson algorithm (by using “maxLik” package in R software [26]) and the ECM algorithm introduced later, are applied. We choose the initial values of p_m, g_0, g_1, λ_1 and λ_2 as follows: $\hat{p}_m^{(0)} = \#(F = 1)/\#(F\in\{0, 1\})$, $\hat{g}_0^{(0)} = \#(M = 0)/\#(M\in\{0, 1, 2\})$, $\hat{g}_1^{(0)} = \#(M = 1)/\#(M\in\{0, 1, 2\})$, $\hat{\lambda}_1^{(0)} = n_{011}/n_{010}$ and $\hat{\lambda}_2^{(0)} = (n_{011} n_{112})/(n_{010} n_{111})$, where # denotes the counting measure. The details about the choice of these initial values are shown in Additional file 1: Appendix B.

MLEs of λ_1 and λ_2 using family trios with parental genotypes missing

It is common that parental genotypes are missing in some family trios. For trios with paternal or maternal genotype missing, we call them “mother-daughter pairs” or “father--daughter pairs”, denoted by *MC* and *FC*, respectively. Thus, *MC* takes possible genotypes from Ω_{MC}={00, 01, 10, 11, 12, 21, 22}, and *FC* takes possible genotypes from Ω_{FC}={00, 01, 11, 12}. As for trios with both parental genotypes missing, we refer to them as “single daughters”. The probabilities of the mother-daughter pair *MC*, father-daughter pair *FC* and single daughter *C* given the daughter being affected are respectively $P(MC|D) = \sum_{F\in\{0,1\}} P(FMC|D)$, $P(FC|D) = \sum_{M\in\{0,1,2\}} P(FMC|D)$ and $P(C|D) = \sum_{F\in\{0,1\}} \sum_{M\in\{0,1,2\}} P(FMC|D)$, where $P(FMC|\, D)$ are given in Table 1.

Suppose that we collect n_{FMC} family trios of type *FMC*, $n_{1m,\, MC}$ mother-daughter pairs of type *MC*, $n_{1f,\, FC}$

father-daughter pairs of type *FC* and $n_{0,\,C}$ single daughters of type *C*, where the subscripts $1m$, $1f$ and 0 respectively mean that each trio has only a mother, only a father and no parents. Then, the log-likelihood function of the observed data is

$$\ln L(\theta) = \sum_{FMC\in\Omega} n_{FMC}\ \ln P(FMC|D) + \sum_{MC\in\Omega_{MC}} n_{1m,MC}\ \ln P(MC|D) + \sum_{FC\in\Omega_{FC}} n_{1f,FC}\ \ln P(FC|D) + \sum_{C\in\{0,1,2\}} n_{0,C}\ \ln P(C|D). \tag{4}$$

Let N_2, N_{1m}, N_{1f} and N_0 respectively be the numbers of family trios, mother-daughter pairs, father-daughter pairs and single daughters. Then, $N_2 = \sum_{FMC\in\Omega} n_{FMC}$, $N_{1m} = \sum_{MC\in\Omega_{MC}} n_{1m,MC}, N_{1f} = \sum_{FC\in\Omega_{FC}} n_{1f,FC}$, $N_0 = \sum_{C\in\{0,1,2\}} n_{0,C}$ and the total sample size $N = N_2 + N_{1m} + N_{1f} + N_0$.

Since it is not so easy to obtain the MLE of θ from the above observed log-likelihood function (4), the ECM algorithm will be employed. Assume that $n_{1m,MC} = \sum_{F\in\{0,1\}} z_{1m,FMC}$, $n_{1f,FC} = \sum_{M\in\{0,1,2\}} z_{1f,FMC}$ and $n_{0,\,C}= \sum_{F\in\{0,1\}}\sum_{M\in\{0,1,2\}} z_{0,FMC}$, where $z_{1m,\,FMC}$, $z_{1f,\,FMC}$ and $z_{0,\,FMC}$ are the unobserved numbers of trios *FMC* for mother-daughter pairs *MC*, father-daughter pairs *FC* and single daughters *C*, respectively (see Additional file 1: Tables S1-S3). Then, the log-likelihood function for the complete data (n_{FMC}, $z_{1m,\,FMC}$, $z_{1f,\,FMC}$, $z_{0,\,FMC}$) can be written as

$$\ln L_C(\theta) = \sum_{FMC\in\Omega} (n_{FMC} + z_{1m,FMC} + z_{1f,FMC} + z_{0,FMC}) \ln P(FMC|D).$$

The following ECM algorithm contains one E-step and five CM-steps at each iteration. In the E-step at iteration $(k + 1)$, we obtain the conditional expectation of $\ln L_C(\theta)$ with respect to the conditional distributions of $z_{1m,\,FMC}$, $z_{1f,\,FMC}$ and $z_{0,\,FMC}$ given $n_{1m,\,MC}$, $n_{1f,\,FC}$ and $n_{0,\,C}$, respectively. $z_{1m,\,FMC} \mid n_{1m,\,MC}$, $z_{1f,\,FMC} \mid n_{1f,\,FC}$ and $z_{0,\,FMC} \mid n_{0,\,C}$ follow the binomial distributions with respective success probabilities $P(F|\,MC, D)$, $P(M|\,FC, D)$ and $P(FM|\,C, D)$. Thus, the *Q* function is given by

$$Q\left(\theta|\hat{\theta}^{(k)}\right) = \sum_{FMC\in\Omega} \left[n_{FMC} + E_{\hat{\theta}^{(k)}}(z_{1m,FMC}|n_{1m,MC}) + E_{\hat{\theta}^{(k)}}\left(z_{1f,FMC}|n_{1f,FC}\right) + E_{\hat{\theta}^{(k)}}(z_{0,FMC}|n_{0,C})\right] \ln P(FMC|D), \tag{5}$$

where $\hat{\theta}^{(k)} = (\hat{p}_m^{(k)}, \hat{g}_0^{(k)}, \hat{g}_1^{(k)}, \hat{\lambda}_1^{(k)}, \hat{\lambda}_2^{(k)})^T$ is the MLE of θ at iteration k,

$$E_{\hat{\theta}^{(k)}}(z_{1m,FMC}|n_{1m,MC}) = n_{1m,MC} P\left(F|MC, D; \hat{\theta}^{(k)}\right) = n_{1m,MC} \frac{P\left(FMC|D; \hat{\theta}^{(k)}\right)}{\sum_{F'\in\{0,1\}} P\left(F'MC|D; \hat{\theta}^{(k)}\right)},$$

$$E_{\hat{\theta}^{(k)}}(z_{1f,FMC}|n_{1f,FC}) = n_{1f,FC} P\left(M|FC, D; \hat{\theta}^{(k)}\right) = n_{1f,FC} \frac{P\left(FMC|D; \hat{\theta}^{(k)}\right)}{\sum_{M'\in\{0,1,2\}} P\left(FM'C|D; \hat{\theta}^{(k)}\right)}$$

and

$$E_{\hat{\theta}^{(k)}}(z_{0,FMC}|n_{0,C}) = n_{0,C} P\left(FM|C, D; \hat{\theta}^{(k)}\right) = n_{0,C} \frac{P\left(FMC|D; \hat{\theta}^{(k)}\right)}{\sum_{F'\in\{0,1\}} \sum_{M'\in\{0,1,2\}} P\left(F'M'C|D; \hat{\theta}^{(k)}\right)}.$$

In the CM-steps, the *Q* function is maximized with respect to each of components of θ in turn, with the others fixed at their previous values. The MLE of θ at iteration $(k + 1)$ are given in Additional file 1: Appendix B. The initial value of θ is obtained only based on N_2 complete family trios when $N_2 \neq 0$ (see Additional file 1: Appendix B). However, when $N_2 = 0$, we estimate the initial values of λ_1 and λ_2 by replacing unknown n_{010}, n_{011}, n_{111} and n_{112} values in $\hat{\lambda}_1^{(0)} = n_{011}/n_{010}$ and $\hat{\lambda}_2^{(0)} = (n_{011}n_{112})/(n_{010}n_{111})$ by their respective conditional expectations (see Additional file 1: Tables S1-S3). For example, n_{011} is replaced by

$$E\left(z_{1m,011}|n_{1m,11}\right) + E\left(z_{1f,011}|n_{1f,01}\right) + E\left(z_{0,011}|n_{0,1}\right)$$

$$= n_{1m,11}\hat{q}_m^{(0)} + n_{1f,01}\cdot\frac{0.5\hat{g}_1^{(0)}}{0.5\hat{g}_1^{(0)} + \hat{g}_2^{(0)}} + n_{0,1}\cdot\frac{0.5\hat{q}_m^{(0)}\hat{g}_1^{(0)}}{\hat{p}_m^{(0)}\hat{g}_0^{(0)} + 0.5\hat{g}_1^{(0)} + \hat{q}_m^{(0)}\hat{g}_2^{(0)}},$$

where $\hat{p}_m^{(0)}$, $\hat{q}_m^{(0)} = 1 - \hat{p}_m^{(0)}$, $\hat{g}_0^{(0)}$, $\hat{g}_1^{(0)}$ and $\hat{g}_2^{(0)} = 1 - \hat{g}_0^{(0)} - \hat{g}_1^{(0)}$ are the initial values of p_m, q_m, g_0, g_1 and g_2, respectively. The details about the choice of these initial values are shown in Additional file 1: Appendix B. Given the initial value of θ, the steps mentioned above continue until the convergence criterion is satisfied. For example, the absolute differences between the estimates of the parameters at two consecutive iterations are all less than 10^{-7}. In addition, note that the ECM algorithm still works when there are no missing genotypes in all the family trios. However, it contains only the CM steps in this situation and can be regarded as a special case of the cyclic coordinate ascent method, which is simple and stable [23].

Confidence interval of γ based on likelihood method

To obtain the CI, we first construct a LRT for testing the null hypothesis H_0: $\gamma = \gamma_0$ as follows,

$$\mathrm{LRT} = 2\ \ln \frac{L\left(\hat{\theta}\right)}{L\left(\tilde{\theta}_0\right)},$$

where $\hat{\theta} = (\hat{p}_m, \hat{g}_0, \hat{g}_1, \hat{\lambda}_1, \hat{\lambda}_2)^T$ is the MLE of θ under H_1. Let θ_0= $(p_m, g_0, g_1, \lambda_2)^T$ be the parameter vector under H_0 with $\gamma_0 = 2(\lambda_1 - 1)/(\lambda_2 - 1)$ (i.e., $\lambda_1 = \gamma_0(\lambda_2 - 1)/2 + 1$), and then $\tilde{\theta}_0 = (\tilde{p}_m, \tilde{g}_0, \tilde{g}_1, \tilde{\lambda}_2)^T$ denotes the MLE of θ_0. The choice of the initial value of θ_0 and the solution of $\tilde{\theta}_0$ using family trios with missing parental genotypes is given in Additional file 1: Appendix B. The LRT asymptotically follows a chi-square distribution with the degree of freedom being one (i.e., χ_1^2).

At the significance level α, the 100(1 – α)% confidence interval of γ based on the LRT is

$$\{\gamma : \mathrm{LRT}(\gamma) < \ \chi^2_{1-\alpha,1}\},$$

and the confidence limits are the values that satisfy

$$\mathrm{LRT}(\gamma) = \chi^2_{1-\alpha,1}. \tag{6}$$

Note that there is no closed-form solution of Eq. (6). Thus, numerical method is applied, such as functions from "rootSolve" package in R software [26]. Let γ_L and γ_U be two unequal roots of Eq. (6) with $\gamma_L < \gamma_U$. Generally, the 100(1 – α)% CI of γ would be (γ_L, γ_U). However, since the true value of γ is bounded in [0, 2], the original estimated CI of γ will be truncated by [0, 2] if necessary. As such, the ultimate CI of γ is $(\gamma_L, \gamma_U)\cap[0, 2]$, which is easier to be interpreted than the origin CI.

Discontinuity problem of confidence interval of γ

Note that γ is a ratio, so like other ratio estimates [27], we find that the proposed CI may consist of two disjoint intervals, such as [0, 0.03)∪(0.59, 2]. In this article, this kind of CIs is referred to as "discontinuous CI" for convenience. Let's take a close look at this discontinuity problem by the following example. Consider the situation of $(n_{000}, n_{010}, n_{011}, n_{021}, n_{101}, n_{111}, n_{112}, n_{122})$=(191, 89, 112, 54, 114, 59, 62, 19). Then, $\hat{\gamma}$ is 1.92 and two roots of Eq. (6) are 0.03 and 0.59, respectively. If the CI is set to be (0.03, 0.59) normally, to our surprise, $\hat{\gamma}$ is located outside this CI. When testing the null hypothesis H_0: $\gamma = \gamma_0$, we find that γ_0 taking values between 0.03 and 0.59 is rejected by the LRT. This means that the interval (0.03, 0.59) is actually a rejection region of the corresponding LRT rather than an acceptance region. Hence, the CI of γ turns to be (−∞, 0.03)∪(0.59,+∞), and [0, 0.03)∪(0.59, 2] after being truncated. The discontinuous CI may occur when the denominator of the ratio is close to zero (i.e., λ_2 is close to 1 in this article) [28]. In fact, when $\lambda_2 = 1$, we assume that we cannot obtain information on the XCI skewing pattern according to the CI of γ. This is because our proposed method measures XCI skewness in the presence of association between the disease and genotypes (i.e., $\lambda_2 \neq 1$). On the other hand, although these discontinuous CIs are considered to be uninformative and are difficult to be interpreted, there is no satisfactory "objective" method for dealing with this problem well [27].

Simulation settings

To assess the performance of the proposed method, we conduct the following simulation study. The sample size N is taken to be 700, which is close to that of RA data (757 pedigrees) [29]. We consider six different combinations of (N_2, N_{1m}, N_{1f}, N_0), which are referred to as six "missing patterns" (MP1–MP6) for convenience and are shown in Table 2. When the missing pattern changes from MP1 to MP4, the number of case-parents trios N_2 decreases and the number of single daughters N_0 increases with $N_{1m} = N_{1f}$. For MP5 and MP6, the number of mother-daughter pairs N_{1m} is different from that of father-daughter pairs N_{1f} with $N_2 = N_0$. In addition, (p_m, p_f) is set to be (0.30, 0.30), (0.25, 0.30), (0.30, 0.25), (0.20, 0.20), (0.15, 0.20) and (0.20, 0.15), and we assume that ρ=0 and 0.05, and λ_2=1.5 and 2. λ_1 is calculated from $\lambda_1 = \gamma(\lambda_2 - 1)/2 + 1$, where γ varies from 0 to 2 in increments of 0.5. Given p_m, p_f, ρ, λ_2 and γ, N_2 case-parents trios are randomly generated from a multinomial distribution with probabilities $P(FMC|\,D)$ shown in Table 1. Similarly, N_{1m} mother-daughter pairs, N_{1f} father-daughter pairs and N_0 single daughters are randomly drawn from the multinomial distributions with probabilities $P(MC|D) = \sum_{F\in\{0,1\}} P(FMC|D)$, $P(FC|D) = \sum_{M\in\{0,1,2\}} P(FMC|D)$ and $P(C|D) = \sum_{F\in\{0,1\}} \sum_{M\in\{0,1,2\}} P(FMC|D)$, respectively. The simulations are based on k=10,000 replicates and 5% significance level.

Table 2 Six simulation settings for different combinations of (N_2, N_{1m}, N_{1f}, N_0) with total sample size N being fixed at 700

MP	N_2	N_{1m}	N_{1f}	N_0
1	700	0	0	0
2	350	100	100	150
3	0	200	200	300
4	0	100	100	500
5	100	400	100	100
6	100	100	400	100

We assess the statistical properties of the CI by the following indexes. Let the coverage probability (CP) be the proportion that the CI contains the true value of γ among k replicates. Note that under H_0: $\gamma = \gamma_0$, the estimated type I error rate of the LRT is 1−CP. ML and MR denote the left tail error and the right tail error (missing the true value of γ), respectively, with $\mathrm{ML} = \#[(\gamma < \gamma_L) \cap (\gamma_L \le \hat{\gamma} \le \gamma_U)]/k$ and $\mathrm{MR} = \#[(\gamma > \gamma_U) \cap (\gamma_L \le \hat{\gamma} \le \gamma_U)]/k$. Further, we use ML/(ML + MR) to measure the balance of ML and MR, which will be close to 0.5 when the balance is achieved. Notice that we do not consider those discontinuous CIs when calculating ML and MR, since we cannot distinguish between the left side and the right side of the discontinuous CIs. As such, we use $\mathrm{DP} = 1 - \#(\gamma_L \le \hat{\gamma} \le \gamma_U)/k$ to represent the proportion of the discontinuous CIs among k replicates. In addition, note that the distribution of a ratio is not necessarily symmetric [30, 31], and the median can be always used to estimate the central tendency of a skewed distribution better than the mean [32]. So, we give the median of the point estimates of γ over k replicates under each simulation scenario. Further, for simulating the power of the LRT, we fix γ_0 at 0, 1 and 2. Finally, we also compare the simulation results under MP1 (consisting of only 700 family trios with both parents) based on the ECM algorithm with those based on the Newton-Raphson algorithm. It is found that the results of the two algorithms are almost consistent with each other (data not shown for brevity). Therefore, we only give the simulation results on the basis of the ECM algorithm in the following section.

Results

Simulation results

Table 3 lists the estimated CP, ML/(ML + MR) and DP of the likelihood-based CI of γ against MP and γ with ρ=0, λ_2=1.5, and (p_m, p_f) being (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25). It is shown in the table that the CP is around 95% under the situations considered. On the other hand, we find that ML/(ML + MR) and DP appear not to be greatly affected by (p_m, p_f). However, the value of γ has strong effect on ML/(ML + MR) and DP. When γ takes values on the boundary (i.e., 0 and 2), ML/(ML + MR) always stays close to 1 and 0, respectively, which indicates extreme imbalance of two tail errors. DP increases as γ gets close to the boundary. Also, the missing pattern has great influence on both ML/(ML + MR) and DP. When the missing pattern varies from MP1 to MP4, where the number of case-parents trios decreases and that of single daughters increases, ML/(ML + MR) becomes more and more far away from 0.5 and DP sharply increases. We also find that under MP5, where the number of mother-daughter pairs is larger than that of father-daughter pairs, ML/(ML + MR) is a little closer to 0.5 and DP becomes smaller compared to those under MP6.

Table 4 shows the corresponding statistical properties of the CI of γ with ρ=0, λ_2=2, and (p_m, p_f) being (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25). As expected, the CP is still controlled well, the ML/(ML + MR) is closer to 0.5 and DP is lower with larger λ_2. We also investigate the effect of ρ=0.05, and the corresponding results are similar to those of ρ=0 (see Additional file 1: Tables S4 and S5). On the other hand, when (p_m, p_f) is set to be (0.20, 0.20), (0.15, 0.20) and (0.20, 0.15), the results are similar to those when (p_m, p_f) being taken as (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25) (see Additional file 1: Tables S6–S9). In addition, the median of the point estimates of γ among k replicates under each simulation scenario is shown in Additional file 1: Figures S1 and S2. From Additional file 1: Figure S1, we can see that the median of $\hat{\gamma}$ gets more far away from the true value of γ as the missing pattern varies from MP1 to MP4, and it is always slightly closer to γ under MP5 than that under MP6. The increase of the value of λ_2 also improves the accuracy of the median of $\hat{\gamma}$, while the values of ρ, p_m and p_f seem to have no great influence on the median of $\hat{\gamma}$.

The simulated powers of the LRT for testing $H_0: \gamma = \gamma_0$ with (p_m, p_f) = (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25) are given in Figs. 1, 2, 3, 4. Fig. 1 shows the simulated powers of the LRT against γ under MP1–MP4 with ρ= 0 and λ_2=1.5. From the first row to the third row of the panels in Fig. 1, (p_m, p_f) is set to be (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25), respectively. From the first column to the third column, γ_0 is fixed at 0, 1 and 2, respectively. It is found that the power increases as the value of $|\gamma - \gamma_0|$ gets larger. For example, in Fig. 1(a), when testing for H_0: $\gamma = \gamma_0$ with γ_0= 0 (XCI skewing completely against mutant allele), the power under γ=1.5 (75% cells express mutant allele) is greater than that under γ= 0.5 (25% cells express mutant allele). On the other hand, when the missing pattern changes from MP1 to MP4, the loss in power is always substantial. Also, we compare the corresponding powers under MP5 and MP6 in Fig. 2. The power under MP5 is always higher than that under MP6 when $\gamma \ne \gamma_0$, which implies that the mother-daughter pairs contain more information on the skewness of XCI than the father-daughter pairs. This is not surprising because when the father's genotype is missing in a trio, it can be inferred according to the available mother's and daughter's genotypes, except for the mother-daughter pair of type MC = 11, whereas we cannot infer the missing mother's genotypes from any father-daughter pairs. In addition, Figs. 3 and 4 give the simulated powers of the LRT under ρ=0 and λ_2=2 for MP1–MP4 and MP5–MP6, respectively. It is shown that the increase of λ_2 leads to the growth in power (Fig. 3 vs. Fig. 1, Fig. 4 vs. Fig. 2). We also simulate the

Table 3 Statistical properties of likelihood-based confidence interval of γ against missing pattern (MP) and γ with ρ=0, λ_2=1.5, and (p_m, p_f) being (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25)[a]

		(p_m, p_f) = (0.30, 0.30)			(p_m, p_f) = (0.25, 0.30)			(p_m, p_f) = (0.30, 0.25)		
MP	γ	CP (%)	ML/(ML + MR)	DP	CP (%)	ML/(ML + MR)	DP	CP (%)	ML/(ML + MR)	DP
1	0	94.37	1	0.098	94.46	1	0.099	94.74	1	0.100
	0.5	94.70	0.51	0.034	94.59	0.51	0.030	94.79	0.56	0.035
	1	95.01	0.33	0.018	94.94	0.34	0.022	94.96	0.34	0.023
	1.5	95.10	0.26	0.052	95.08	0.30	0.048	95.08	0.25	0.058
	2	95.13	0	0.061	94.84	0	0.060	94.97	0	0.072
2	0	95.17	1	0.162	94.69	1	0.161	94.78	1	0.160
	0.5	94.98	0.59	0.064	94.87	0.53	0.055	95.21	0.59	0.061
	1	94.96	0.21	0.036	95.05	0.20	0.038	95.21	0.17	0.037
	1.5	94.65	0.16	0.104	95.01	0.12	0.105	94.67	0.09	0.110
	2	94.58	0	0.130	94.98	0	0.120	94.87	0	0.140
3	0	94.79	1	0.373	95.28	1	0.347	94.99	1	0.372
	0.5	95.57	0.75	0.140	95.58	0.62	0.119	95.58	0.75	0.138
	1	95.49	0.01	0.071	95.72	0.02	0.069	95.56	0.03	0.063
	1.5	94.84	0	0.232	94.93	0	0.227	94.96	0	0.210
	2	94.87	0	0.413	94.75	0	0.385	95.13	0	0.393
4	0	94.72	1	0.500	94.59	1	0.472	94.76	1	0.488
	0.5	95.14	0.87	0.183	95.31	0.81	0.162	95.32	0.89	0.174
	1	94.99	0.05	0.067	95.13	0.03	0.073	94.85	0.04	0.070
	1.5	94.79	0	0.229	94.86	0	0.238	94.65	0	0.194
	2	94.65	0	0.457	94.39	0	0.460	94.81	0	0.402
5	0	94.77	1	0.211	95.12	1	0.204	95.08	1	0.209
	0.5	95.56	0.67	0.077	95.61	0.56	0.066	95.60	0.66	0.078
	1	95.04	0.12	0.050	95.06	0.13	0.052	95.49	0.11	0.047
	1.5	94.72	0.04	0.155	94.83	0.07	0.150	94.69	0.05	0.150
	2	94.74	0	0.220	94.66	0	0.199	94.81	0	0.231
6	0	95.08	1	0.295	95.25	1	0.275	95.05	1	0.291
	0.5	95.28	0.72	0.108	95.43	0.62	0.094	95.68	0.77	0.106
	1	95.20	0.06	0.059	95.30	0.05	0.063	95.03	0.03	0.059
	1.5	94.57	0.01	0.190	95.09	0.03	0.183	94.64	0.01	0.186
	2	94.87	0	0.294	94.69	0	0.267	94.64	0	0.297

[a] The simulations are conducted under 10,000 replicates and 5% significance level

powers under ρ=0.05 and (p_m, p_f) = (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25), which are similar to those under ρ=0 (see Additional file 1: Figures S3–S6). Finally, the simulated powers under (p_m, p_f) = (0.20, 0.20), (0.15, 0.20) and (0.20, 0.15) are given in Additional file 1: Figures S7–S14, which are always lower than those under (p_m, p_f) = (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25), respectively.

Application to RA data

Rheumatoid arthritis (RA) is an autoimmune disease, which has been reported to be associated with the skewness of XCI [33]. To investigate the XCI skewing patterns at the X-linked loci associated with RA, we apply our proposed method to the data from North American Rheumatoid Arthritis Consortium [29], which is made available from Genetic Analysis Workshop 15 [34]. The dataset includes 757 pedigrees and 293 single nucleotide polymorphism (SNP) markers on X chromosome. In this application, one nuclear family with a typed affected daughter is selected randomly from each pedigree. As such, a total of 703 nuclear families are included, which contains 64 case-parents trios, 179 mother-daughter pairs, 37 father-daughter pairs and 423 single daughters.

Since our proposed method is applicable in the presence of association, we ultimately measure the degree of XCI skewing at five SNPs which have been found to be associated

Table 4 Statistical properties of likelihood-based confidence interval of γ against missing pattern (MP) and γ with ρ=0, λ_2=2, and (p_m, p_f) being (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25)[a]

		(p_m, p_f) = (0.30, 0.30)			(p_m, p_f) = (0.25, 0.30)			(p_m, p_f) = (0.30, 0.25)		
MP	γ	CP (%)	ML/(ML + MR)	DP	CP (%)	ML/(ML + MR)	DP	CP (%)	ML/(ML + MR)	DP
1	0	94.67	0.88	0.025	94.73	0.90	0.030	94.80	0.93	0.029
	0.5	94.71	0.40	0.005	94.95	0.42	0.006	94.65	0.41	0.008
	1	94.86	0.41	0.004	94.97	0.40	0.004	95.09	0.41	0.006
	1.5	94.96	0.41	0.006	95.25	0.43	0.005	95.11	0.41	0.008
	2	95.14	0.02	0.022	94.99	0.01	0.024	95.08	0.01	0.024
2	0	95.17	1	0.032	94.71	0.99	0.040	94.64	1	0.038
	0.5	95.16	0.40	0.025	95.13	0.35	0.022	95.00	0.38	0.027
	1	94.60	0.36	0.019	95.24	0.38	0.019	95.01	0.33	0.020
	1.5	94.62	0.39	0.026	95.32	0.43	0.020	94.93	0.40	0.034
	2	94.71	0	0.028	94.81	0	0.026	95.13	0	0.029
3	0	94.91	1	0.168	95.10	1	0.189	95.09	1	0.172
	0.5	95.24	0.29	0.159	95.61	0.27	0.135	95.56	0.31	0.164
	1	95.58	0.02	0.097	95.51	0.02	0.101	95.25	0	0.091
	1.5	95.21	0.03	0.280	95.06	0.04	0.250	94.92	0	0.296
	2	95.16	0	0.176	95.07	0	0.142	95.10	0	0.219
4	0	94.74	1	0.387	94.68	1	0.406	94.57	1	0.393
	0.5	94.80	0.43	0.263	94.61	0.32	0.227	94.51	0.46	0.275
	1	95.19	0.01	0.127	95.37	0	0.155	95.20	0.01	0.110
	1.5	94.67	0	0.449	94.54	0	0.455	94.69	0	0.405
	2	94.65	0	0.433	94.71	0	0.399	94.61	0	0.462
5	0	94.75	1	0.050	94.94	1	0.063	94.93	1	0.054
	0.5	95.18	0.36	0.050	95.73	0.37	0.046	95.31	0.39	0.058
	1	94.98	0.25	0.037	94.85	0.26	0.042	95.21	0.22	0.037
	1.5	94.97	0.23	0.084	94.71	0.31	0.065	94.85	0.25	0.093
	2	95.24	0	0.037	94.42	0	0.035	94.63	0	0.051
6	0	95.34	1	0.083	95.10	1	0.095	95.20	1	0.101
	0.5	95.72	0.36	0.092	95.24	0.36	0.081	95.48	0.33	0.094
	1	94.93	0.09	0.059	94.91	0.12	0.062	94.83	0.10	0.063
	1.5	94.96	0.14	0.133	94.57	0.18	0.119	95.07	0.13	0.146
	2	94.81	0	0.055	94.71	0	0.054	94.95	0	0.075

[a] The simulations are conducted under 10,000 replicates and 5% significance level

with RA by the XMCPDT method at the significance level of 1% [35]. Notice that the XMCPDT method is conducted based on 246 pedigrees from the RA dataset. We identify the at-risk allele by the value of $\hat{\lambda}_2$, and denote the estimates of the frequencies of the at-risk allele in males and females obtained from the ECM algorithm by $\hat{p}_m$ and $\hat{p}_f$, respectively. Table 5 lists the p-value of XMCPDT, the values of ($\hat{p}_m$, $\hat{p}_f$), $\hat{\lambda}_2$ and $\hat{\gamma}$, and 95% CI of γ based on the proposed likelihood method for each of five SNPs. From Table 5, we find that there are three SNPs (rs916685, rs1043034 and rs2005463) with the 95% CIs containing the value of γ=1, which indicates the random XCI. On the other hand, the XCI skewing at rs2238907 is found with $\hat{\gamma}$ =0.35 and the 95% CI being [0, 0.79), which suggests that the skewness of XCI is against the at-risk allele with 17.5% (0.35/2) cells in heterozygous females having the at-risk allele active, while the other 82.5% cells keeping the normal allele active. However, the 95% CI of γ at rs1264064 is [0, 2], providing no information on the XCI skewing pattern. In addition, we evaluate $\hat{\gamma}$'s and the 95% CIs of γ at the rest 288 SNPs in the RA dataset, and find that there are 21 SNPs with nonrandom XCI pattern. But note that, if we assume that all of 293 SNPs except for rs2238907 are under

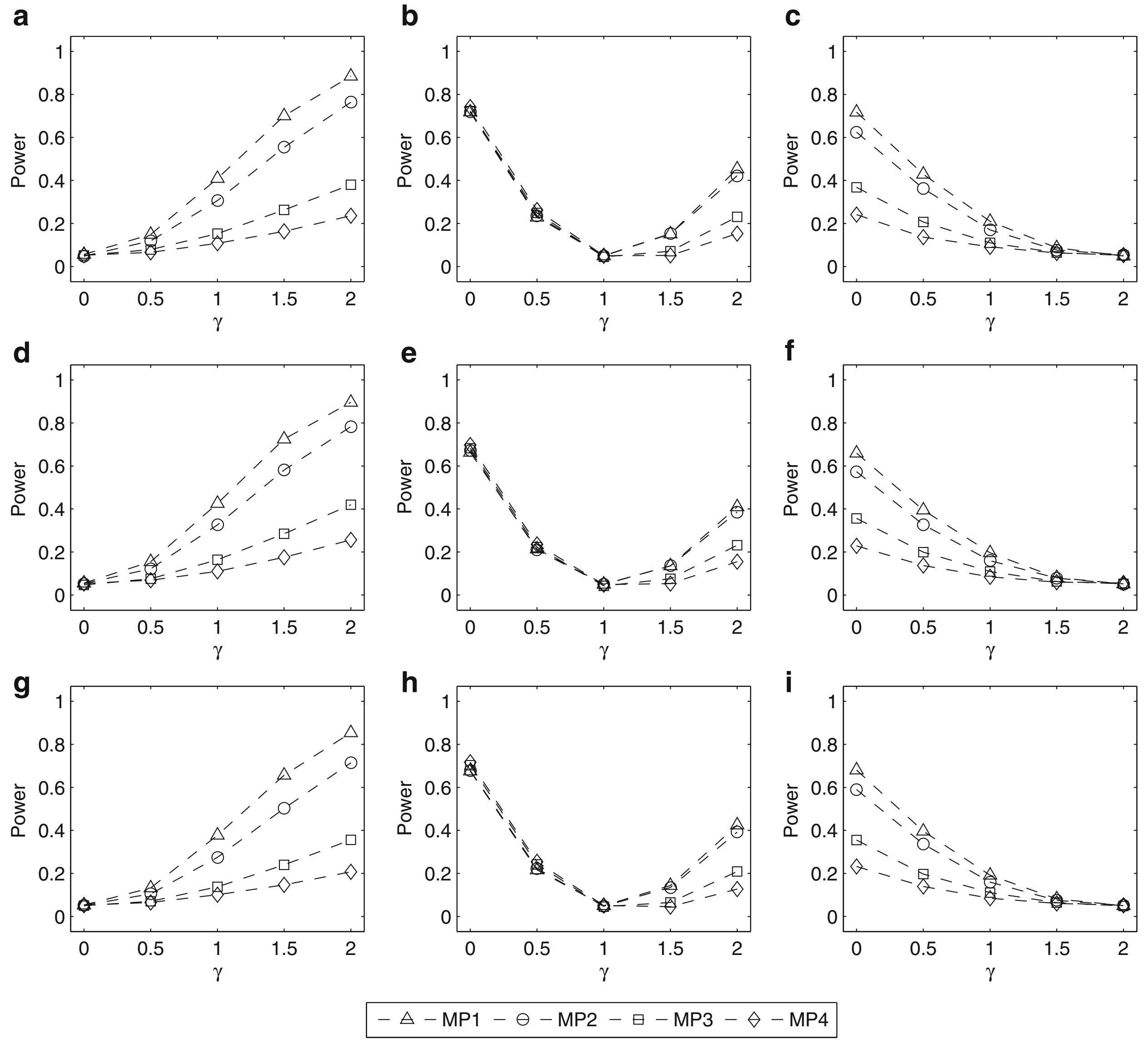

Fig. 1 Estimated powers of LRT against γ under MP1–MP4 with ρ=0 and λ_2= 1.5. The results are based on 10,000 replicates and 5% significance level. **a** $(p_m, p_f) = (0.30, 0.30)$ and γ_0=0; **b** $(p_m, p_f) = (0.30, 0.30)$ and γ_0=1; **c** $(p_m, p_f) = (0.30, 0.30)$ and γ_0=2; **d** $(p_m, p_f) = (0.25, 0.30)$ and γ_0=0; **e** $(p_m, p_f) = (0.25, 0.30)$ and γ_0=1; **f** $(p_m, p_f) = (0.25, 0.30)$ and γ_0=2; **g** $(p_m, p_f) = (0.30, 0.25)$ and γ_0=0; **h** $(p_m, p_f) = (0.30, 0.25)$ and γ_0=1; **i** $(p_m, p_f) = (0.30, 0.25)$ and γ_0=2. Note that γ_0=0, 1 and 2 represent XCI skewing completely against mutant allele, random XCI and XCI skewing completely toward mutant allele, respectively

H_0: $\gamma = 1$, then the corresponding false positive rate would be 0.0719 (21/292), which is still below the upper bound $0.05 + 1.96 \times \sqrt{0.05 \times (1{-}0.05)/292} = 0.0750$. Besides, association between these 21 SNPs and RA has not been found by XMCPDT at the 1% significance level, so we should draw conclusions with this caution.

Discussion

In this article, we propose a statistical measure γ of the degree of the XCI skewing for family trio data, which can be represented as a ratio of two GRRs in females in the presence of association between the disease and genotypes. Further, we obtain the point estimate of γ, which is constructed by the MLEs of two GRRs in females. When there are missing parental genotypes in some family trios, the ECM algorithm is used to estimate the two GRRs. The CI of γ is derived from the likelihood method by inverting the LRT. We conduct the simulation study under various simulation settings, including six missing patterns of families, six groups of allele frequencies, two different values of inbreeding coefficient in females, two different values of λ_2 and five different values of γ. The simulation results show that the proposed likelihood-based CI of γ has an accurate CP under

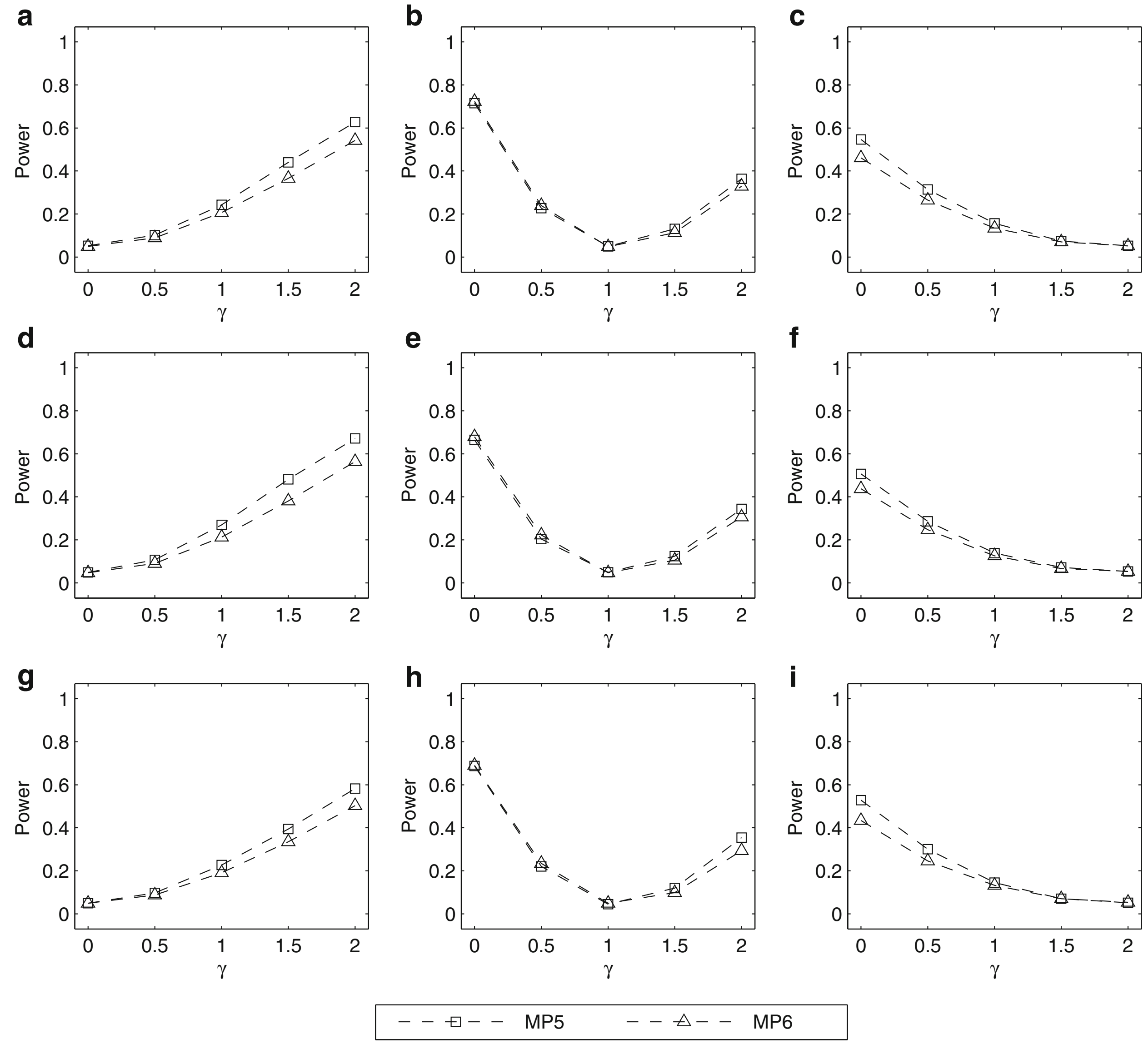

Fig. 2 Estimated powers of LRT against γ under MP5 and MP6 with ρ=0 and λ_2= 1.5. The results are based on 10,000 replicates and 5% significance level. **a** (p_m, p_f) = (0.30, 0.30) and γ_0=0; **b** (p_m, p_f) = (0.30, 0.30) and γ_0=1; **c** (p_m, p_f) = (0.30, 0.30) and γ_0=2; **d** (p_m, p_f) = (0.25, 0.30) and γ_0=0; **e** (p_m, p_f) = (0.25, 0.30) and γ_0=1; **f** (p_m, p_f) = (0.25, 0.30) and γ_0=2; **g** (p_m, p_f) = (0.30, 0.25) and γ_0=0; **h** (p_m, p_f) = (0.30, 0.25) and γ_0=1; **i** (p_m, p_f) = (0.30, 0.25) and γ_0=2. Note that γ_0=0, 1 and 2 represent XCI skewing completely against mutant allele, random XCI and XCI skewing completely toward mutant allele, respectively

the situations considered, while ML/(ML + MR) and DP of the CI of γ and the median of estimates of γ are influenced by the values of γ, λ_2 and the missing pattern. Similarly, the simulated power of LRT is affected by the values of $|\gamma - \gamma_0|$, λ_2, (p_m, p_f) and the missing pattern. Finally, we apply our proposed method to the RA data from USA and find out a locus, rs2238907, which may undergo the XCI skewing against the at-risk allele.

Many X-linked diseases are always associated with XCI skewing in females. Our proposed statistical measure γ provides information on the potential loci subject to XCI skewing, thus it is helpful to uncover the pathogenesis of X-linked diseases. However, most of the statistical studies on X chromosome today focus mainly on the association tests [17, 24, 25, 36–38], so there are no other statistical methods available to measure the skewness of XCI. On the other hand, although the XCI skewing pattern can also be detected by differential methylation between the active and inactive X chromosomes or mRNA transcription in cells, our proposed statistical method takes use of population data to measure the skewness of XCI. Thus, it can reflect the average level of the XCI skewing in female population.

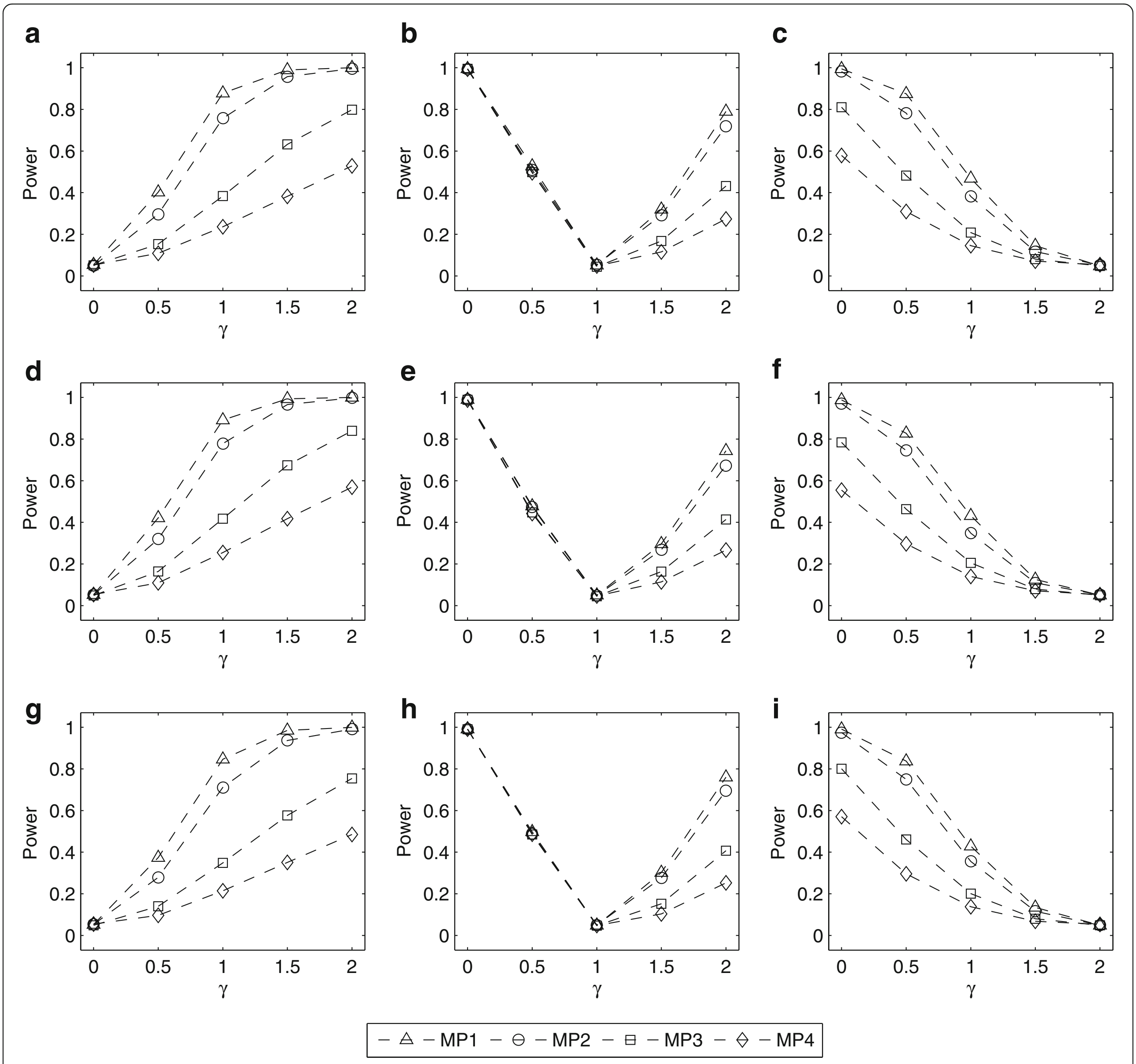

Fig. 3 Estimated powers of LRT against γ under MP1–MP4 with ρ=0 and λ_2=2. The results are based on 10,000 replicates and 5% significance level. **a** (p_m, p_f) = (0.30, 0.30) and γ_0=0; **b** (p_m, p_f) = (0.30, 0.30) and γ_0=1; **c** (p_m, p_f) = (0.30, 0.30) and γ_0=2; **d** (p_m, p_f) = (0.25, 0.30) and γ_0=0; **e** (p_m, p_f) = (0.25, 0.30) and γ_0=1; **f** (p_m, p_f) = (0.25, 0.30) and γ_0=2; **g** (p_m, p_f) = (0.30, 0.25) and γ_0=0; **h** (p_m, p_f) = (0.30, 0.25) and γ_0=1; **i** (p_m, p_f) = (0.30, 0.25) and γ_0=2. Note that γ_0=0, 1 and 2 represent XCI skewing completely against mutant allele, random XCI and XCI skewing completely toward mutant allele, respectively

There are some issues in our proposed method. First of all, the original CI is truncated by [0, 2] to enhance the interpretability of the CI. However, when the whole original CI lies outside [0, 2], the CI ultimately obtained after truncation would be empty. Although it is hard to interpret this kind of CI containing no values, the simulation results show that when γ takes values on the boundary (i.e., 0 and 2), these empty CIs seldom occur. For example, when γ = 0, two tail errors (ML and MR) are extremely imbalance with ML/(ML + MR) being 1 or close to 1. This means that there are no or very few original CIs whose upper limit is below 0. Likewise, ML/(ML + MR) is 0 or close to 0 when γ = 2, which implies that there are no or very few original CIs whose lower limit is beyond 2. On the other hand, the proposed likelihood method has its own drawback in deriving the CI of a ratio like any other ratio estimation methods. We find that the likelihood-based CI of γ may consist of two disjoint intervals, such as $[0, 0.03) \cup (0.59, 2]$, and it is also difficult for us to interpret. For example, if $\hat{\gamma} = 1.92$ and the CI of γ is $[0, 0.03) \cup (0.59, 2]$, then the

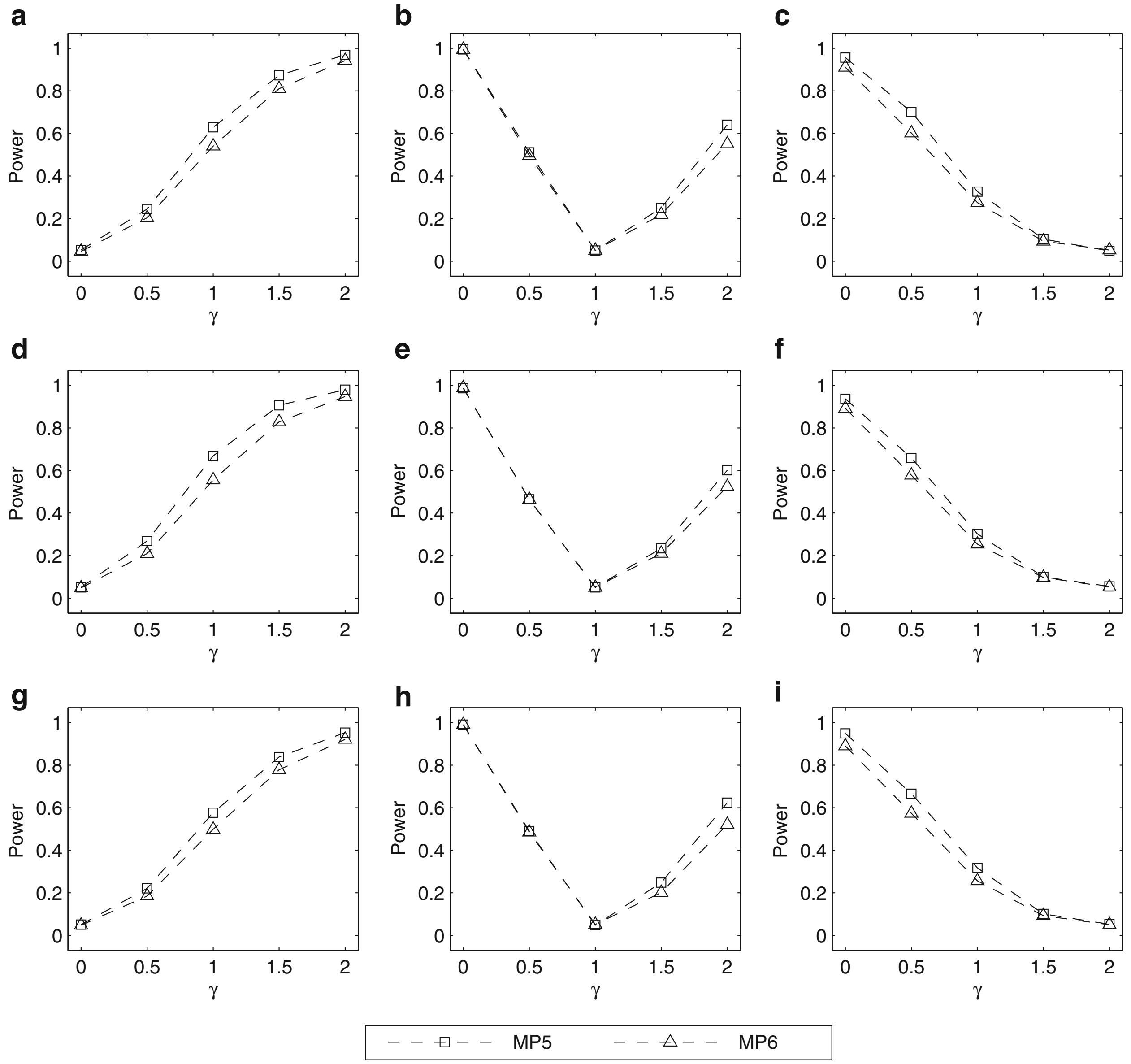

Fig. 4 Estimated powers of LRT against γ under MP5 and MP6 with ρ=0 and λ_2=2. The results are based on 10,000 replicates and 5% significance level. **a** (p_m, p_f) = (0.30, 0.30) and γ_0=0; **b** (p_m, p_f) = (0.30, 0.30) and γ_0=1; **c** (p_m, p_f) = (0.30, 0.30) and γ_0=2; **d** (p_m, p_f) = (0.25, 0.30) and γ_0=0; **e** (p_m, p_f) = (0.25, 0.30) and γ_0=1; **f** (p_m, p_f) = (0.25, 0.30) and γ_0=2; **g** (p_m, p_f) = (0.30, 0.25) and γ_0=0; **h** (p_m, p_f) = (0.30, 0.25) and γ_0=1; **i** (p_m, p_f) = (0.30, 0.25) and γ_0=2. Note that γ_0=0, 1 and 2 represent XCI skewing completely against mutant allele, random XCI and XCI skewing completely toward mutant allele, respectively

Table 5 Application of proposed method to RA dataset with *p*-values of XMCPDT less than 1% significance level

SNP name	*p*-value[a]	$(\hat{p}_m, \hat{p}_f)$	$\hat{\lambda}_2$	$\hat{\gamma}$	95% CI of γ
rs2238907	0.004	(0.20, 0.24)	2.18	0.35	[0, 0.79)
rs916685	0.003	(0.17, 0.20)	2.55	0.53	[0, 1.33)
rs1264064	0.001	(0.42, 0.45)	1.96	0.71	[0, 2]
rs1043034	0.007	(0.19, 0.24)	3.73	0.81	(0.51, 1.69)
rs2005463	0.007	(0.18, 0.23)	4.59	0.61	(0.40, 1.06)

[a] *P*-value of XMCPDT for testing association between SNP and RA [35]

corresponding LRT would reject the null hypothesis of $\gamma_0 = 0.5$, and accept that of $\gamma_0 = 0.01$. It is hard to explain that the LRT rejects a γ_0 being close to $\hat{\gamma}$, while accepts one being far away from $\hat{\gamma}$. Although this kind of CI is undesirable, it is also inevitable and can be regarded as a hint of λ_2 being close to 1 [39]. In addition, it should be noted that the ECM algorithm is not applicable when all the family trios are "single daughters", since the MLE of θ may not be uniquely specified under this situation (the details see Additional file 1: Appendix C). However, if the other family trios were collected, then the single daughters can make contribution to the MLE of θ in the ECM algorithm together with these trio data (the details see Additional file 1: Appendix D). Finally, we assume that the genotypes' frequencies in males and females (p_m, g_0 and g_1) are unknown and estimate them together with λ_1 and λ_2 in our simulation study and real data application. Alternatively, if we can obtain information on the allele frequencies from the online databases, such as the Allele Frequency Net Database [40] and the UCSC Genome Browser Database [41], then it is unnecessary to re-estimate p_m, g_0 and g_1, which will reduce the number of parameters so that the ECM algorithm runs faster.

Note that the ECM algorithm can converge to a local maximum of the log-likelihood function instead of a global maximum [23]. To investigate this, we randomly choose 1000 initial values of θ (θ_0) from the parameter space and regard the MLE of θ (θ_0) with the maximum log-likelihood among 1000 $\ln L(\hat{\theta})$'s ($\ln L(\tilde{\theta}_0)$'s) as the global MLE of θ (θ_0). We conduct a simulation study under the simulation settings with ρ=0, λ_2= 1.5 and (p_m, p_f) = (0.30, 0.30), and the details see Additional file 1: Appendix E. The simulation results (see Additional file 1: Tables S10 and S11) show that the values of $\hat{\theta}$ and $\ln L(\hat{\theta})$ ($\tilde{\theta}_0$ and $\ln L(\tilde{\theta}_0)$) based on one initial value estimated by the method described in Additional file 1: Appendix B are very close to those based on 1000 initial values under all the simulated situations when N_2 (the number of complete family trios) is not too small, such as MP1 and MP2, which may indicate that the ECM algorithm based on the estimated initial value converges towards the global maximum. As for MP3-MP6, except that $\tilde{\theta}_0$ with $(\gamma_0, \gamma) = (1, 2)$ under MP5 and MP6, $\tilde{\theta}_0$ with $(\gamma_0, \gamma) = (1, 1)$ and $(1, 2)$ under MP3, and $\tilde{\theta}_0$ with (γ_0, γ) $= (1, 1)$, $(1, 1.5)$ and $(1, 2)$ under MP4 may converge to a local maximum, all the other $\hat{\theta}$ and $\tilde{\theta}_0$ results converge to the global maximum. Further, for these seven cases, we try and randomly select ten groups of initial values of θ_0 from the parameter space and regard $\tilde{\theta}_0$ with the maximum log-likelihood among ten $\ln L(\tilde{\theta}_0)$'s as the final MLE of θ_0. We find that $\tilde{\theta}_0$'s based on ten and 1000 initial values are very close to each other under all the seven simulated situations (see Additional file 1: Table S11). So, if N_2 is zero or too small, we recommend using multiple initial values (such as ten) for obtaining the global MLE of θ_0. On the other hand, family trios with both parents are always fortunately collected in the family-based studies in practice.

In future studies, we will extend our proposed method to incorporate covariates by using nuclear families with affected and unaffected offspring. Furthermore, to facilitate the interpretability of the CI of γ, we will utilize the prior information, such as the order of the GRRs in females and the information of the presence of association.

Conclusions

The proposed statistical measure for the skewness of XCI is applicable for complete family trio data or family trio data with some paternal genotypes missing. The likelihood-based CI has an accurate CP under the situations considered. Therefore, our proposed statistical measure is generally recommended in practice for discovering the potential loci which undergo the XCI skewing.

Additional file

Additional file 1: Appendix A. Derivation of $P(FMC|D)$ in Table 1. **Appendix B.** Choice of initial value of θ (θ_0) and MLE of θ (θ_0) using family trios with missing parental genotypes. **Appendix C.** Inapplicability of ECM algorithm when using only single daughters. **Appendix D.** Contribution of single daughters to estimate of θ in ECM algorithm. **Appendix E.** Effect of different initial values of θ (θ_0) on ECM algorithm. **Tables S1**–**S3.** The conditional probabilities and conditional expectations for seven types of possible mother-daughter pairs, four types of possible father-daughter pairs and three types of possible single daughters, respectively**. Tables S4**–**S5.** Statistical properties of likelihood-based confidence interval of γ against missing pattern (MP) and γ with ρ=0.05, (p_m, p_f) = (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25), and λ_2=1.5 and 2, respectively. **Tables S6**–**S9.** Statistical properties of likelihood-based confidence interval of γ against missing pattern (MP) and γ with (p_m, p_f)= (0.20, 0.20), (0.15, 0.20) and (0.20, 0.15), ρ=0 and 0.05, and λ_2=1.5 and 2, respectively. **Table S10.** Averages of absolute differences of each element of $\hat{\theta}$ and $\ln L(\hat{\theta})$ between ECM$_1$ and ECM$_{1000}$ with ρ=0, λ_2= 1.5 and (p_m, p_f) = (0.30, 0.30) under MP1-MP6. **Table S11.** Averages of absolute differences of each element of $\tilde{\theta}_0$ and $\ln L(\tilde{\theta}_0)$ between ECM$_1$/ECM$_{10}$ and ECM$_{1000}$ with ρ=0, λ_2= 1.5 and (p_m, p_f) = (0.30, 0.30) under MP1-MP6. **Figures S1**–**S2.** Medians of point estimates of γ against MP for different p_m, p_f and λ_2 values with ρ=0 and 0.05, respectively. **Figures S3**–**S6.** Estimated powers of LRT against γ with ρ=0.05 and (p_m, p_f) being (0.30, 0.30), (0.25, 0.30) and (0.30, 0.25) under MP1–MP4 and MP5–MP6, and λ_2=1.5 and 2, respectively. **Figures S7**–**S14.** Estimated powers of LRT against γ with (p_m, p_f) being (0.20, 0.20), (0.15, 0.20) and (0.20, 0.15) under MP1–MP4 and MP5–MP6, ρ=0 and 0.05, and λ_2=1.5 and 2, respectively. (PDF 2328 kb)

Abbreviations

CI: Confidence interval; CP: Coverage probability; DP: Proportion of the discontinuous CIs; ECM: Expectation-conditional- maximization; GRR: Genotypic relative risk; LRT: Likelihood ratio test; ML: Left tail error; MLE: Maximum likelihood estimate; MP: Missing patterns; MR: Right tail error; RA: Rheumatoid arthritis; SNP: Single nucleotide polymorphism; XCI: X chromosome inactivation; XMCPDT: Monte Carlo pedigree disequilibrium test on X chromosome

Acknowledgments

The authors thank the editor and three reviewers for helpful comments that greatly improve the presentation of the article.

Funding

This study was supported by the National Natural Science Foundation of China grants 81,773,544, 81,373,098 and 81,573,207 and the National and Guangzhou University Students' Innovation and Enterprise Training Project of China grant 201,612,121,017. The authors thank the Genetic Analysis Workshops for providing the RA data, which were supported by the National Institutes of Health grant R01 GM031575. The RA data were gathered with the support of grants from the National Institutes of Health grants N01-AR-2-2263 and R01-AR-44422, and the National Arthritis Foundation.

Availability of data and materials

The dataset supporting the conclusions of this article is from North American Rheumatoid Arthritis Consortium, which is made available from Genetic Analysis Workshop 15 (http://www.gaworkshop.org/).

Authors' contributions

SQX, YZ, PW, WL, XBW and JYZ all contributed to the study design, analytical preparation and the writing of the manuscript. SQX and YZ performed the simulation studies. PW, WL, XBW and JYZ analyzed the data and revised the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 12 June 2018 Accepted: 8 November 2018
Published online: 05 December 2018

References

1. Postma DS. Gender differences in asthma development and progression. Gend Med 2007;4 Suppl B: S133–46.
2. Invernizzi P, Pasini S, Selmi C, Gershwin ME, Podda M. Female predominance and X chromosome defects in autoimmune diseases. J Autoimmun. 2009;33:12–6.
3. Dorak MT, Karpuzoglu E. Gender differences in cancer susceptibility: an inadequately addressed issue. Front Genet. 2012;3:268.
4. Qureshi IA, Mehler MF. Genetic and epigenetic underpinnings of sex differences in the brain and in neurological and psychiatric disease susceptibility. Prog Brain Res. 2010;186:77–95.
5. Skuse DH. Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. Pediatr Res. 2000;47:9–16.
6. Chow JC, Yen Z, Ziesche SM, Brown CJ. Silencing of the mammalian X chromosome. Annu Rev Genomics Hum Genet. 2005;6:69–92.
7. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005;434:400–4.
8. Minks J, Robinson WP, Brown CJ. A skewed view of X chromosome inactivation. J Clin Invest. 2008;118:20–3.
9. Van den Veyver IB. Skewed X inactivation in X-linked disorders. Semin Reprod Med. 2001;19:183–91.
10. Brown CJ. Skewed X-chromosome inactivation: cause or consequence? J Natl Cancer Inst. 1999;91:304–5.
11. Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. Skewed X-chromosome is a common feature of X-linked mental retardation disorders. Am J Hum Genet. 2002;71:168–73.
12. Prchal JT, Carroll AJ, Prchal JF, Crist WM, Skalka HW, Gealy WJ, Harley J, Malluh A. Wiskott-Aldrich syndrome: cellular impairments and their implication for carrier detection. Blood. 1980;56:1048–54.
13. Conley ME, Lavoie A, Briggs C, Brown P, Guerra C, Puck JM. Nonrandom X chromosome inactivation in B cells from carriers of X chromosome-linked severe combined immunodeficiency. Proc Natl Acad Sci U S A. 1988;85:3090–4.
14. Salsano E, Tabano S, Sirchia SM, Colapietro P, Castellotti B, Gellera C, Rimoldi M, Pensato V, Mariotti C, Pareyson D, Miozzo M, Uziel G. Preferential expression of mutant ABCD1 allele is common in adrenoleukodystrophy female carriers but unrelated to clinical symptoms. Orphanet J Rare Dis. 2012;7:10.
15. Kristiansen M, Langerød A, Knudsen GP, Weber BL, Børresen-Dale AL, ørstavik KH. High frequency of skewed X inactivation in young breast cancer patients. J Med Genet. 2002;39:30–3.
16. Clayton D. Testing for association on the X chromosome. Biostatistics. 2008;9:593–600.
17. 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;38:483–93.
18. Kubota T. A new assay for the analysis of X-chromosome inactivation in carriers with an X-linked disease. Brain Dev. 2001;23(Suppl 1):S177–81.
19. Busque L, Zhu J, DeHart D, Griffith B, Willman C, Carroll R, Black PM, Gilliland DG. An expression based clonality assay at the human androgen receptor locus (HUMARA) on chromosome X. Nucleic Acids Res. 1994;22:697–8.
20. Szelinger S, Malenica I, Corneveaux JJ, Siniard AL, Kurdoglu AA, Ramsey KM, Schrauwen I, Trent JM, Narayanan V, Huentelman MJ, Craig DW. Characterization of X chromosome inactivation using integrated analysis of whole-exome and mRNA sequencing. PLoS One. 2014;9:e113036.
21. Carrel L, Willard HF. Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others. Proc Natl Acad Sci U S A. 1999;96:7364–9.
22. Cotton AM, Lam L, Affleck JG, Wilson IM, Peñaherrera MS, McFadden DE, Kobor MS, Lam WL, Robinson WP, Brown CJ. Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. Hum Genet. 2011;130:187–201.
23. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika. 1993;80:267–78.
24. Chen Z, Ng HKT, Li J, Liu Q, Huang H. Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. Stat Methods Med Res. 2017;26:567–82.
25. Wang P, Xu SQ, Wang BQ, Fung WK, Zhou JY. A robust and powerful test for case–control genetic association study on X chromosome. Stat Methods Med Res. 2018. https://doi.org/10.1177/0962280218799532.
26. Team RC. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2013. http://www.r-project.org. 2018.
27. Liseo B. Bayesian and conditional frequentist analyses of the Fieller's problem. A critical review Metron 2003;LXI:133–50.
28. Hwang JTG. Fieller's problems and resampling techniques. Stat Sinica. 1995; 5:161–71.
29. Amos CI, Chen WV, Remmers E, Siminovitch KA, Seldin MF, Criswell LA, Lee AT, John S, Shephard ND, Worthington J, Cornelis F, Plenge RM, Begovich AB, Dyer TD, Kastner DL, Gregersen PK. Data for genetic analysis workshop (GAW) 15 problem 2, genetic causes of rheumatoid arthritis and associated traits. BMC Proc. 2007;1(Suppl 1):S3.
30. Marsaglia G. Ratios of normal variables and ratios of sums of uniform variables. J Am Stat Assoc. 1965;60:193–204.
31. Diaz-Francés E, Rubio FJ. On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. Stat Pap. 2013;54:309–23.
32. Wilcox RR, Keselman HJ. Modern robust data analysis methods: measures of central tendency. Psychol Methods. 2003;8:254–74.
33. Chabchoub G, Uz E, Maalej A, Mustafa CA, Rebai A, Mnif M, Bahloul Z, Farid NR, Ozcelik T, Ayadi H. Analysis of skewed X-chromosome inactivation in females with rheumatoid arthritis and autoimmune thyroid diseases. Arthritis Res Ther. 2009;11:R106.
34. Genetic Analysis Workshop. 1982. https://www.gaworkshop.org. Accessed 3 Oct 2018.
35. Zou QL, You XP, Li JL, Fung WK, Zhou JY. A powerful parent-of-origin effects test for qualitative traits on X chromosome in general pedigrees. BMC Bioinformatics. 2018;19:8.
36. Hickey PF, Bahlo M. X chromosome association testing in genome wide association studies. Genet Epidemiol. 2011;35:664–70.
37. Ma L, Hoffman G, Keinan A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. BMC Genomics. 2015;16:241.
38. Choi S, Lee S, Qiao D, Hardin M, Cho MH, Silverman EK, Park T, Won S. FARVAT**X**: family-based rare variant association test for X-linked genes. Genet Epidemiol. 2016;40:475–85.
39. Gleser LJ, Hwang JT. The nonexistence of 100(1-α)% confidence sets of finite expected diameter in errors-in-variables and related models. Ann Stat. 1987;15:1351–62.
40. Allele Frequency Net Database. 2015. http://www.allelefrequencies.net/. Accessed 3 Oct 2018.
41. UCSC Genome browser database. 2000. http://genome.ucsc.edu/. Accessed 3 Oct 2018.