PLOS ONE

RESEARCH ARTICLE

Likelihood Ratio Test for Excess Homozygosity at Marker Loci on X Chromosome

Xiao-Ping You, Qi-Lei Zou, Jian-Long Li, Ji-Yuan Zhou*

State Key Laboratory of Organ Failure Research and Guangdong Provincial Key Laboratory of Tropical Research, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong, China

* zhoujiyuan5460@hotmail.com

OPEN ACCESS

Citation: You X-P, Zou Q-L, Li J-L, Zhou J-Y (2015) Likelihood Ratio Test for Excess Homozygosity at Marker Loci on X Chromosome. PLoS ONE 10(12): e0145032. doi:10.1371/journal.pone.0145032

Editor: Dmitri Zaykin, NIH - National Institute of Environmental Health Sciences, UNITED STATES

Received: March 11, 2015

Accepted: November 28, 2015

Published: December 15, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The rheumatoid arthritis data are from North American Rheumatoid Arthritis Consortium, which are available from Genetic Analysis Workshop 15.

Funding: This work was supported by the National Natural Science Foundation of China (81373098), Science and Technology Planning Project of Guangdong Province, China (2013B021800038) to Professor Ji-Yuan Zhou. The Genetic Analysis Workshops are 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 (NO1-AR-2-2263 and RO1-AR-

Abstract

The assumption of Hardy-Weinberg equilibrium (HWE) is generally required for association analysis using case-control design on autosomes; otherwise, the size may be inflated. There has been an increasing interest of exploring the association between diseases and markers on X chromosome and the effect of the departure from HWE on association analysis on X chromosome. Note that there are two hypotheses of interest regarding the X chromosome: (i) the frequencies of the same allele at a locus in males and females are equal and (ii) the inbreeding coefficient in females is zero (without excess homozygosity). Thus, excess homozygosity and significantly different minor allele frequencies between males and females are used to filter X-linked variants. There are two existing methods to test for (i) and (ii), respectively. However, their size and powers have not been studied yet. Further, there is no existing method to simultaneously detect both hypotheses till now. Therefore, in this article, we propose a novel likelihood ratio test for both (i) and (ii) on X chromosome. To further investigate the underlying reason why the null hypothesis is statistically rejected, we also develop two likelihood ratio tests for detecting (i) and (ii), respectively. Moreover, we explore the effect of population stratification on the proposed tests. From our simulation study, the size of the test for (i) is close to the nominal significance level. However, the size of the excess homozygosity test and the test for both (i) and (ii) is conservative. So, we propose parametric bootstrap techniques to evaluate their validity and performance. Simulation results show that the proposed methods with bootstrap techniques control the size well under the respective null hypothesis. Power comparison demonstrates that the methods with bootstrap techniques are more powerful than those without bootstrap procedure and the existing methods. The application of the proposed methods to a rheumatoid arthritis dataset indicates their utility.

44422), and the National Arthritis Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Association analysis is a useful tool to map disease loci by using markers on autosomes based on family data and case-control data [1–9]. There has been an increasing interest of exploring the association between diseases and markers on X chromosome and the effect of the departure from Hardy-Weinberg equilibrium (HWE) on association analysis on X chromosome [10–17]. Note that there are two hypotheses of interest regarding the X chromosome: (i) the frequencies of the same allele at a locus in males and females are equal and (ii) the inbreeding coefficient in females is zero (without excess homozygosity) in X-specific quality control [18, 19]. As such, excess homozygosity in females and significantly different minor allele frequencies between males and females are used to filter X-linked variants [20, 21]. The inbreeding coefficient is generally estimated by functions of excess homozygosity [22, 23], which may be caused by population substructure, consanguineous mating or factors like null alleles [24, 25]. Overall and Nichols developed an approach to distinguish population substructure and consanguinity by using multilocus genotype data [24]. On the other hand, Zheng et al. proposed two test statistics to test for the equality of the frequencies of the same allele in males and females and the zero inbreeding coefficient in females on X chromosome, respectively [14]. However, they only focused on association analysis on X chromosome and the type I error rates and powers of these two test statistics have not been studied yet. Further, there is no existing method to simultaneously detect both of the issues till now.

Therefore, in this article, we first combine two test statistics proposed in zheng et al. [14] and suggest Z_0 to simultaneously test for (i) the equality of the frequencies of the same allele in males and females and (ii) the zero inbreeding coefficient on X chromosome based on the collected sample. For the purpose of improving the test power for both (i) and (ii), a novel likelihood ratio test on X chromosome is proposed. We write out the likelihood functions of the collected sample under the null hypothesis and alternative hypothesis at a single locus on X chromosome, respectively. Next, we obtain the maximum likelihood estimates (MLEs) of the unknown parameters by expectation-maximization (EM) algorithms [26] and construct the corresponding likelihood ratio test (LRT_0) statistic to test for both (i) and (ii). If the null hypothesis is statistically rejected, we further conduct two hypothesis testing issues to find the underlying reasons why the null hypothesis is violated by proposing another two likelihood ratio tests LRT_1 (for the equality of the frequencies of the same allele in males and females) and LRT_2 (for excess homozygosity). Note that the size of LRT_0 and LRT_2 is conservative from our simulation study. As such, we use parametric bootstrap techniques to evaluate the validity and performance of LRT_0 and LRT_2, which are respectively denoted by LRT_{0b} and LRT_{2b}. Moreover, we explore the effect of population stratification on the proposed tests. In addition, the root mean squared error (RMSE) and bias are used to assess the accuracy of the MLEs of the unknown parameters. Finally, the application of the proposed methods to a rheumatoid arthritis (RA) dataset indicates their utility.

Materials and Methods

Background and notations

Consider a biallelic marker locus on X chromosome with alleles M_1 and M_2. Let p_m and p_f be the frequencies of M_1 in males and females, respectively. As such, the frequencies of M_2 in males and females are $q_m = 1 - p_m$ and $q_f = 1 - p_f$, respectively. In females, let ρ be the inbreeding coefficient, which is generally nonnegative [27–29]. Thus, the frequencies of three genotypes M_1M_1, M_1M_2 and M_2M_2 in females can be expressed as follows:

$$P(M_1M_1) = p_f^2 + \rho p_f q_f, P(M_1M_2) = 2(1-\rho)p_f q_f, P(M_2M_2) = q_f^2 + \rho p_f q_f.$$

To this end, there is no excess homozygosity in females when $\rho = 0$; excess homozygosity exists when $\rho > 0$. Note that $p_m \neq p_f$ may be true on X chromosome. So, we construct the null hypothesis denoted by H_0: $p_m = p_f$ and $\rho = 0$ to test for both of the hypotheses (i) and (ii). If the null hypothesis is violated, we need to investigate which one of $p_m \neq p_f$ and $\rho > 0$ is true. As such, we have other two hypothesis testing issues with the null hypothesis being H_{01}: $p_m = p_f$ and H_{02}: $\rho = 0$, respectively. It should be noted that X chromosome has the problem of X chromosome inactivation and dosage compensation [30], but we do not consider them in this section. The corresponding discussion can be found later (see the Discussion section).

Assume that n_{1m} and n_{0m} represent the numbers of males with alleles M_1 and M_2 in a collected sample, respectively; n_{2f}, n_{1f} and n_{0f} denote the numbers of females with genotypes M_1M_1, M_1M_2 and M_2M_2, respectively. Then, $N_m = n_{1m} + n_{0m}$ and $N_f = n_{2f} + n_{1f} + n_{0f}$ are respectively the numbers of males and females in the sample, and $N = N_m + N_f$ is the sample size.

Existing methods Z_1 and Z_2 for H_{01} (equality of the frequencies of the same allele in males and females) and H_{02} (zero inbreeding coefficient), respectively

Zheng et al. [14] proposed the test statistic

$$Z_1 = \frac{(\hat{p}_m - \hat{p}_f)^2}{Var(\hat{p}_m) + Var(\hat{p}_f)}$$

to test for H_{01}: $p_m = p_f$, where $\hat{p}_m = n_{1m}/N_m$ and $\hat{p}_f = (2n_{2f} + n_{1f})/(2N_f)$ are the estimates of p_m and p_f, $Var(\hat{p}_m) = \hat{p}_m(1 - \hat{p}_m)/N_m$, and $Var(\hat{p}_f) = [\hat{p}_f - 2\hat{p}_f^2 + \hat{P}(M_1M_1)]/(2N_f)$ are the estimates of the variances of $\hat{p}_m$ and $\hat{p}_f$ under H_{01}, respectively, with $\hat{P}(M_1M_1) = n_{2f}/N_f$. Under H_{01}, Z_1 asymptotically follows the chi-square distribution with one degree of freedom when the sample size is large enough.

Weir and cockerham [31] introduced the disequilibrium coefficient in females $\Delta_f = P(M_1M_1) - p_f^2 = \rho p_f q_f$. In other words, testing for $\Delta_f = 0$ is equivalent to testing for $\rho = 0$. Hence, zheng et al. [14] further developed the following test statistic to test for H_{02}: $\rho = 0$,

$$Z_2 = \frac{[\hat{\Delta}_f + \hat{p}_f\hat{q}_f/(2N_f)]^2}{Var(\hat{\Delta}_f)} = N_f\frac{[\hat{\Delta}_f + \hat{p}_f\hat{q}_f/(2N_f)]^2}{\hat{p}_f{}^2\hat{q}_f{}^2},$$

where $\hat{\Delta}_f = \hat{P}(M_1M_1) - \hat{p}_f^2$, $\hat{q}_f = 1 - \hat{p}_f$, $E(\hat{\Delta}_f) = -p_f q_f/(2N_f)$ and $Var(\hat{\Delta}_f) = p_f^2 q_f{}^2/N_f$. Under H_{02}, Z_2 approximately follows the chi-square distribution with one degree of freedom when N_f is large enough. It should be noted that the test Z_2 has nothing to do with male individuals and thus only needs female individuals.

Z_0 test for both hypotheses (i) and (ii) of interest regarding the X chromosome

Zheng et al. [14] showed that, under H_0: $p_m = p_f$ and $\rho = 0$, Z_1 and Z_2 are independent. However, they did not propose the corresponding test statistic for H_0. As such, we suggest the test statistic

$$Z_0 = Z_1 + Z_2$$

to test for H_0: $p_m = p_f$ and $\rho = 0$. Under H_0, Z_0 asymptotically follows the chi-square distribution

with the degrees of freedom being 2. Moreover, it should be noted that we can use

$$\hat{\rho}_z = \frac{\hat{P}(M_1M_1) - \hat{p}_f^2}{\hat{p}_f\hat{q}_f}$$

to estimate the inbreeding coefficient ρ.

Likelihood ratio test for both hypotheses (i) and (ii) of interest regarding the X chromosome

To construct a likelihood ratio test (LRT) for H_0: $p_m = p_f$ and $\rho = 0$, we give the likelihood function of the sample as follows:

$$\begin{aligned} L(\theta) &= \binom{N_m}{n_{1m}} \binom{N_f}{n_{2f}, n_{1f}, n_{0f}} p_m^{n_{1m}} q_m^{n_{0m}} (p_f^2 + \rho p_f q_f)^{n_{2f}} \\ &\quad \times [2(1-\rho)p_f q_f]^{n_{1f}} (q_f^2 + \rho p_f q_f)^{n_{0f}}, \end{aligned} \tag{1}$$

where $\theta = (p_m, p_f, \rho)$. Firstly, we use the following EM algorithm to estimate the unknown parameters p_m, p_f and ρ under the alternative hypothesis (H_1: $p_m \neq p_f$ or $\rho > 0$). Suppose that $Y = (Y_1, Y_2, Y_3, Y_4, Y_5) = (n_{1m}, n_{0m}, n_{2f}, n_{1f}, n_{0f})$ denotes the observed data. $(Y_1, Y_2, Y_3, Y_4, Y_5)$ can be augmented by splitting the third cell into two cells W_1 and W_2, which are unobservable random variables such that $Y_3 = W_1 + W_2$ for female homozygote M_1M_1 and W_1 and W_2 follow the binomial distributions with success probabilities $p_f^2/(p_f^2 + \rho p_f q_f)$ and $\rho p_f q_f/(p_f^2 + \rho p_f q_f)$, respectively, and by splitting the fifth cell into two cells W_3 and W_4, where $Y_5 = W_3 + W_4$ for female homozygote M_2M_2 and W_3 and W_4 follow the binomial distributions with success probabilities $q_f^2/(q_f^2 + \rho p_f q_f)$ and $\rho p_f q_f/(q_f^2 + \rho p_f q_f)$, respectively. Thus, the likelihood function of complete data $(n_{1m}, n_{0m}, w_1, w_2, n_{1f}, w_3, w_4)$ is:

$$L_c(\theta) \propto p_m^{n_{1m}} q_m^{n_{0m}} p_f^{2w_1+w_2+n_{1f}+w_4} q_f^{w_2+n_{1f}+2w_3+w_4} \rho^{w_2+w_4} (1-\rho)^{n_{1f}},$$

where the normalizing constant is omitted for brevity.

At the E-step, the Q function at iteration $(k+1)$ is constructed as

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= n_{1m}\ln p_m + n_{0m}\ln q_m \\ &\quad + [2E_{\theta^{(k)}}(w_1|n_{2f}) + E_{\theta^{(k)}}(w_2|n_{2f}) + n_{1f} + E_{\theta^{(k)}}(w_4|n_{0f})]\ln p_f \\ &\quad + [E_{\theta^{(k)}}(w_2|n_{2f}) + n_{1f} + 2E_{\theta^{(k)}}(w_3|n_{0f}) + E_{\theta^{(k)}}(w_4|n_{0f})]\ln(1-p_f) \\ &\quad + [E_{\theta^{(k)}}(w_2|n_{2f}) + E_{\theta^{(k)}}(w_4|n_{0f})]\ln\rho + n_{1f}\ln(1-\rho), \end{aligned}$$

where $\theta^{(k)}$ is the estimate of θ at iteration k.

At the M-step, the estimated value $\theta^{(k+1)}$ of θ at iteration $(k+1)$ can be obtained by maximizing the Q function with respect to θ. Therefore, the MLEs of p_m, p_f and ρ at iteration $(k+1)$

are respectively

$$\begin{aligned}\hat{p}_m &= \frac{n_{1m}}{N_m},\\ \hat{p}_{f1}^{(k+1)} &= \frac{E_{\theta^{(k)}}(2w_1 + w_2|n_{2f}) + n_{1f} + E_{\theta^{(k)}}(w_4|n_{0f})}{2N_f},\\ \hat{\rho}_1^{(k+1)} &= \frac{E_{\theta^{(k)}}(w_2|n_{2f}) + E_{\theta^{(k)}}(w_4|n_{0f})}{E_{\theta^{(k)}}(w_2|n_{2f}) + E_{\theta^{(k)}}(w_4|n_{0f}) + n_{1f}}.\end{aligned}$$

Note that the MLE of p_m is the same for all the iterations, which is also the same as zheng et al. [14]. In the above expressions,

$$E_{\theta^{(k)}}(w_1|n_{2f}) = \frac{n_{2f}\left(\hat{p}_{f1}^{(k)}\right)^2}{\left(\hat{p}_{f1}^{(k)}\right)^2 + \hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}, \quad (2)$$

$$E_{\theta^{(k)}}(w_2|n_{2f}) = \frac{n_{2f}\hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}{\left(\hat{p}_{f1}^{(k)}\right)^2 + \hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}, \quad (3)$$

$$E_{\theta^{(k)}}(w_3|n_{0f}) = \frac{n_{0f}\left(\hat{q}_{f1}^{(k)}\right)^2}{\left(\hat{q}_{f1}^{(k)}\right)^2 + \hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}, \quad (4)$$

$$E_{\theta^{(k)}}(w_4|n_{0f}) = \frac{n_{0f}\hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}{\left(\hat{q}_{f1}^{(k)}\right)^2 + \hat{\rho}_1^{(k)}\hat{p}_{f1}^{(k)}\hat{q}_{f1}^{(k)}}, \quad (5)$$

where $\hat{q}_{f1}^{(k)} = 1 - \hat{p}_{f1}^{(k)}$. Given the initial value $\theta^{(0)}$ of θ, the above-mentioned two steps 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}. The value of θ obtained at the last iteration is taken as the MLE $\hat{\theta}_1 = (\hat{p}_m, \hat{p}_{f1}, \hat{\rho}_1)$ of θ under H_1.

Note that $p_m = p_f$ and $\rho = 0$ under H_0. Let $p = p_m = p_f$, the pooled allele frequency of M_1. Then, $L(\theta)$ in Eq (1) can be rewritten as

$$L(\theta) \propto p^{n_{1m}+2n_{2f}+n_{1f}}(1-p)^{n_{0m}+n_{1f}+2n_{0f}}.$$

Thus, the MLE of p under H_0 is $\hat{p} = (n_{1m} + 2n_{2f} + n_{1f})/(N_m + 2N_f)$, the estimated pooled allele frequency of M_1. Let $\hat{\theta}_0 = (\hat{p}, \hat{p}, 0)$. Then, we can construct the following LRT to test for H_0

$$\mathrm{LRT}_0 = 2\ln\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_0)}, \quad (6)$$

which asymptotically follows a chi-square distribution with the degrees of freedom being 2 when the null hypothesis holds.

Likelihood ratio test for equality of frequencies of the same allele in males and females

Once the null hypothesis (H_0: $p_m = p_f$ and $\rho = 0$) is rejected based on the result of Eq (6), we further need to consider the following two tests H_{01}: $p_m = p_f$ and H_{02}: $\rho = 0$. Note that under the null hypothesis H_{01}: $p_m = p_f = p$, ρ may not be zero and we need to estimate it. Let $\phi = (p, \rho)$ and $q = 1 - p$. Thus, the corresponding likelihood function of complete data is

$$L_{c_1}(\phi) \propto p^{n_{1m}+2w_1+w_2+n_{1f}+w_4}(1-p)^{n_{0m}+w_2+n_{1f}+2w_3+w_4}\rho^{w_2+w_4}(1-\rho)^{n_{1f}}.$$

We use the following EM algorithm to estimate ϕ under H_{01}. The corresponding formulas at iteration $(k + 1)$ are as follows

$$\hat{p}_{01}^{(k+1)} = \frac{E_{\phi^{(k)}}(2w_1 + w_2|n_{2f}) + E_{\phi^{(k)}}(w_4|n_{0f}) + n_{1f} + n_{1m}}{N_m + 2N_f},$$

$$\hat{\rho}_{01}^{(k+1)} = \frac{E_{\phi^{(k)}}(w_2|n_{2f}) + E_{\phi^{(k)}}(w_4|n_{0f})}{E_{\phi^{(k)}}(w_2|n_{2f}) + E_{\phi^{(k)}}(w_4|n_{0f}) + n_{1f}},$$

where $\hat{p}_{01}^{(k+1)}$ and $\hat{\rho}_{01}^{(k+1)}$ are respectively the MLEs of p and ρ at iteration $(k + 1)$, and $\phi^{(k)} = (\hat{p}_{01}^{(k)}, \hat{\rho}_{01}^{(k)})$. $E_{\phi^{(k)}}(w_1|n_{2f})$, $E_{\phi^{(k)}}(w_2|n_{2f})$, $E_{\phi^{(k)}}(w_3|n_{0f})$ and $E_{\phi^{(k)}}(w_4|n_{0f})$ in the above expressions are similar to $E_{\theta^{(k)}}(w_1|n_{2f})$, $E_{\theta^{(k)}}(w_2|n_{2f})$, $E_{\theta^{(k)}}(w_3|n_{0f})$ and $E_{\theta^{(k)}}(w_4|n_{0f})$ in Eqs (2)–(5), just replacing $\hat{p}_{f1}^{(k)}$, $\hat{q}_{f1}^{(k)}$ and $\hat{\rho}_1^{(k)}$ in Eqs (2)–(5) by $\hat{p}_{01}^{(k)}$, $\hat{q}_{01}^{(k)}$ and $\hat{\rho}_{01}^{(k)}$, respectively. Let $\hat{\theta}_{01} = (\hat{p}_{01}, \hat{p}_{01}, \hat{\rho}_{01})$. Then, we propose the following test statistic LRT_1 to test for the null hypothesis H_{01}: $p_m = p_f$,

$$\mathrm{LRT}_1 = 2\ln\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_{01})}, \tag{7}$$

which approximately follows a chi-square distribution with the degree of freedom being 1 under H_{01}.

Likelihood ratio test for inbreeding coefficient being zero

Note that under the null hypothesis $H_{02}: \rho = 0$, p_m and p_f may be different from each other and we need to estimate them separately. Let $\psi = (p_m, p_f)$ and $L(\theta)$ in Eq (1) can be rewritten as

$$L_2(\psi) \propto p_m^{n_{1m}} q_m^{n_{0m}} p_f^{2n_{2f}+n_{1f}} q_f^{n_{1f}+2n_{0f}}.$$

Then, the MLEs of p_m and p_f are $\hat{p}_m = n_{1m}/N_m$ and $\hat{p}_f = (2n_{2f} + n_{1f})/(2N_f)$, respectively, which are the same as zheng et al. [14]. Let $\hat{\theta}_{02} = (\hat{p}_m, \hat{p}_f, 0)$. As such, we develop the following test statistic to test for $H_{02}: \rho = 0$

$$\mathrm{LRT}_2 = 2\ln\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_{02})}, \tag{8}$$

which asymptotically follows a chi-square distribution with the degree of freedom being 1 under H_{02}. Just like the Z_2 test statistic, LRT_2 only uses female individuals in the sample because the terms based on male individuals in the numerator and the denominator of the fraction are the same, which can be reduced.

Likelihood ratio tests via parametric bootstrap for H_0 and H_{02}

It should be noted from our simulation results (see the Results section) that the simulated type I error rates of LRT_0 and LRT_{02} respectively for H_0 and H_{02} are too conservative. On the other hand, several studies showed that the likelihood ratio tests may typically not follow a chi-square distribution asymptotically [31, 32], and hence their exact distributions can be obtained by Monte Carlo simulation [33]. Accordingly, we make use of parametric bootstrap techniques to evaluate the size and power of these two methods. For convenience, we denote these methods via parametric bootstrap by LRT_{0b} and LRT_{2b}, respectively. We begin by describing the implementation steps for LRT_{0b} as follows:

1. For a collected sample of size N with N_m males and N_f females, calculate the value of LRT_0;
2. Compute the estimated pooled allele frequency $\hat{p}$ based on the sample as follows: $\hat{p} = (n_{1m} + 2n_{2f} + n_{1f})/(N_m + 2N_f)$;
3. Based on $\hat{p}$, calculate the frequencies of three genotypes M_1M_1, M_1M_2 and M_2M_2 in females under H_0 in the following: $\hat{p}^2$, $2\hat{p}\hat{q}$ and $\hat{q}^2$, respectively, where $\hat{q} = 1 - \hat{p}$;
4. According to $\hat{p}$ and $\hat{q}$, regenerate the alleles of N_m males; based on $\hat{p}^2$, $2\hat{p}\hat{q}$ and $\hat{q}^2$, regenerate the genotypes of N_f females;
5. Calculate the value of LRT_0 based on the new N_m males and N_f females, denoted by LRT_0^*;
6. Repeat Steps 4 and 5 B times, which results in B test statistics LRT_0^{1*}, LRT_0^{2*}, . . ., LRT_0^{B*};
7. The P-value of the original LRT_0 can be estimated as

$$\hat{P} - \text{value} = \frac{1}{B}\sum_{i=1}^{B} I_{\{LRT_0^{i*} > \text{LRT}_0\}}.$$

For LRT_{2b}, we can conduct the steps similar to those mentioned above. Firstly, after obtaining the value of LRT_2, calculate the frequencies of three genotypes M_1M_1, M_1M_2 and M_2M_2 in females under H_{02} in the following: $\hat{p}_f^2$, $2\hat{p}_f\hat{q}_f$ and $\hat{q}_f^2$, respectively, with $\hat{p}_f = (2n_{2f} + n_{1f})/(2N_f)$. The alleles of N_m males stay the same as the original sample and only regenerate the genotypes of N_f females according to $\hat{p}_f^2$, $2\hat{p}_f\hat{q}_f$ and $\hat{q}_f^2$. Then, carry out the similar procedures of Steps 4–7 and we can obtain the the estimated P-value of LRT_2.

Software implementation

We have written the XHWE software with R (http://www.r-project.org), which includes the eight test statistics: LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2. The R package named XHWE is available on CRAN (http://cran.r-project.org/web/packages/XHWE/). The initial values of p_m, p_f, p and ρ in the EM algorithms are taken to be n_{1m}/N_m, $(2n_{2f} + n_{1f})/(2N_f)$, $(n_{1m} + 2n_{2f} + n_{1f})/(N_m + 2N_f)$ and 0.02, respectively. The convergence criterion is that the absolute differences between the estimates of the parameters at two consecutive iterations are all less than 10^{-7} for the LRT-type statistics. The default maximum number of iterations is 1000. The input data file is the standard pedigree data. The XHWE software only uses the founders with genotypes available in it and will analyze marker loci one by one. The software outputs the values of all the test statistics and the corresponding P-values. Also, the XHWE software outputs the estimates of all the parameters under both the null and alternative hypotheses for each test statistic. The parameter estimates under the alternative hypothesis for the LRT-type test

statistics are the same. However, under the respective null hypotheses of the LRT-type test statistics, the estimates may be different. It should be noted that the estimates of p_m and p_f under the null hypothesis of H_{02} in this article and those in zheng et al. [14] are the same, respectively. The output results will be automatically saved in the text file named "results.txt".

Simulation settings

Simulation study is conducted to assess the performance of the proposed LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b} and Z_0 test statistics and to compare them with the existing Z_1 and Z_2 under various simulation settings which are similar to those in zheng et al. [14]. The allele frequency p_m in males takes two values: 0.3 and 0.5. When p_m is fixed, the value of p_f in females is taken as $p_f = p_m + \epsilon$, where $\epsilon = 0, \pm 0.04$ and ± 0.05. The inbreeding coefficient ρ in females is set at 0 to 0.1 in increment of 0.05. The sample size is taken as 800 and 1200 with the ratio $r = N_m : N_f$ being 2:1, 1.5:1, 1:1, 1:1.5 and 1:2. As mentioned earlier, when $p_m = p_f$ and $\rho = 0$, the size of all the eight test statistics is simulated; when $p_m = p_f$ and $\rho > 0$, the size of LRT_1 and Z_1 is gotten; when $p_m \neq p_f$ and $\rho = 0$, the size of LRT_2, LRT_{2b} and Z_2 is obtained. Otherwise, we simulate the corresponding powers. In addition, it should be noted that for the fixed sample size (800 or 1200) simulated above, the powers of all the three test statistics LRT_2, LRT_{2b} and Z_2 for $H_{02} : \rho = 0$ are not so large, from our simulation results below. On the other hand, these three test statistics only use female individuals. As such, we further obtain the sample size N_f required for LRT_{2b} to gain 80% simulated power and then simulate the size and powers of LRT_2, LRT_{2b} and Z_2 under this sample size. To investigate how population structure affects the proposed methods, we also consider the following population stratification model with two subpopulations in our simulation study. $p_m = 0.3$ (0.5), $p_f = p_m + \epsilon$, $\epsilon = 0, \pm 0.04$ and ± 0.05 in the first (second) subpopulation and the ϵ values are respectively denoted by ϵ_1 and ϵ_2. Assume that $\rho = 0$ in each subpopulation, and the ratio of each subpopulation constructing the population is set to 0.5. The sample size is taken to be 1800, where each individual is a female or a male with equal probability. Note that under population stratification, the null hypothesis H_0: $p_m = p_f$ and $\rho = 0$ is generally not true. Thus, we use the population stratification model to study the powers of the proposed methods. The significance level is fixed at 5% and 10000 replications are simulated under each simulation setting. For LRT_{0b} and LRT_{2b} via parametric bootstrap, B is set to be 1000. Finally, to compare the efficiency of the parameter estimates of the proposed EM algorithms with those in zheng et al. [14] for each simulation setting, we use the RMSEs and biases to assess the accuracy of the parameter estimates, where $\text{RMSE} = \sqrt{[\text{Bias}(\hat{\beta})]^2 + Var(\hat{\beta})}$ and $\text{Bias} = E(\hat{\beta}) - \beta$, and β is the parameter which needs to estimate.

Results

Simulation results

Table 1 lists the simulated size of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 under $H_0 : p_m = p_f = p$ and $\rho = 0$ with $N = 800$ and 1200 and $p = 0.3$ and 0.5 for different values of $r = N_m : N_f$. According to the table, the size of LRT_1, Z_0, Z_1 and Z_2 is close to the nominal 5% level, while the size of LRT_0 and LRT_2 is too conservative. However, after the parametric bootstrap technique, LRT_{0b} and LRT_{2b} stay close to the nominal 5% level.

Fig 1 gives the simulated powers of the eight test statistics against r under $H_1 : p_m \neq p_f$ and $\rho > 0$ for different values of ρ (0.05 and 0.1) and N (800 and 1200), having $p_m = 0.3$ and $p_f = 0.35$. It is shown in the figure that LRT_{0b} is more powerful than LRT_0 and Z_0, and LRT_0 and Z_0 have the similar performance in power (Fig 1a-1d in the first row), regarded of the inbreeding coefficient ρ, the sample size N and the ratio r. LRT_1 and Z_1 have almost the same performance in

Table 1. Simulated size (in %) of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 under $H_0 : p_m = p_f = p$ and $\rho = 0$ with $N = 800$ and 1200 for different values of r and p.

N	r	p	LRT_0	LRT_{0b}	LRT_1	LRT_2	LRT_{2b}	Z_0	Z_1	Z_2
800	2:1	0.3	3.01	4.81	5.02	1.91	4.87	4.83	5.22	4.83
	2:1	0.5	2.98	4.97	5.02	2.28	4.99	5.13	5.19	5.13
	1.5:1	0.3	2.92	4.81	4.74	2.19	4.75	4.99	4.93	4.99
	1.5:1	0.5	3.07	4.93	4.83	2.22	4.98	5.02	4.99	5.02
	1:1	0.3	3.17	5.05	4.74	2.59	5.36	4.82	4.81	4.82
	1:1	0.5	3.30	4.99	5.33	2.40	5.20	5.16	5.34	5.16
	1:1.5	0.3	3.05	5.18	5.09	2.40	5.11	5.09	5.28	5.09
	1:1.5	0.5	3.39	5.18	5.03	2.49	5.34	5.19	5.07	5.19
	1:2	0.3	3.13	4.89	4.77	2.18	4.81	5.13	4.89	5.13
	1:2	0.5	3.12	4.85	4.65	2.32	5.13	5.23	4.78	5.23
1200	2:1	0.3	3.42	5.31	4.84	2.35	5.07	5.47	4.97	5.47
	2:1	0.5	3.45	5.38	4.76	2.48	5.15	5.38	5.01	5.38
	1.5:1	0.3	2.91	4.84	4.84	2.30	5.12	4.83	5.02	5.16
	1.5:1	0.5	3.36	5.31	5.29	2.45	5.35	5.38	5.42	5.38
	1:1	0.3	2.88	4.77	5.05	2.15	4.73	4.73	5.18	4.73
	1:1	0.5	3.04	5.07	5.22	2.01	4.79	4.94	5.30	4.94
	1:1.5	0.3	2.93	4.97	4.75	2.25	4.98	4.98	4.87	4.98
	1:1.5	0.5	3.08	4.83	5.06	2.24	4.86	4.87	5.12	4.87
	1:2	0.3	3.13	4.98	4.83	2.39	5.31	5.03	4.92	5.03
	1:2	0.5	3.24	5.06	4.86	2.54	5.38	5.05	4.91	5.05

doi:10.1371/journal.pone.0145032.t001

power (Fig 1e-1h in the second row). LRT_{2b} has much more power than LRT_2 and Z_2, and LRT_2 is a little less powerful than Z_2 (Fig 1i-1l in the third row). The powers of LRT_1 and Z_1 are not so affected by the different values of r, while LRT_0, LRT_{0b}, Z_0, LRT_{2b}, LRT_2 and Z_2 are more and more powerful with the number of female individuals increasing (r changing from 2:1 to 1:2) when other parameters are fixed. We also find that the powers of LRT_0, LRT_{0b}, Z_0, LRT_2, LRT_{2b} and Z_2 appear great reaction to the different values of ρ when N is fixed. Specially, their powers under $\rho = 0.1$ (subplots in the second and fourth columns, respectively) are much larger than those under $\rho = 0.05$ (subplots in the first and third columns, respectively). However, the powers of LRT_1 and Z_1 are almost not influenced by ρ. Further, it can be seen in Fig 1 that LRT_0, LRT_{0b} and Z_0 with two degrees of freedom (subplots in the first row) are much more powerful than LRT_1, Z_1, LRT_2, LRT_{2b} and Z_2 with one degree of freedom (subplots in the second and third rows). This is because the true model is $p_m \neq p_f$ and $\rho > 0$. In addition, when the sample size changes from 800 (subplots in the first and second columns) to 1200 (subplots in the third and fourth columns), all the test statistics are much more powerful.

Fig 2 displays the simulated size/powers of the eight test statistics against r under $H_{02} : \rho = 0$ for different values of p_f, having $p_m = 0.3$ and $N = 1200$. The results in the third row of the figure are the size of LRT_2, LRT_{2b} and Z_2, while those in the first and the second rows of the figure are the powers of LRT_0, LRT_{0b} and Z_0, and those of LRT_1 and Z_1, respectively. It is shown in the figure that the size of LRT_{2b} and Z_2 maintains close to the nominal 5% level, while LRT_2 is too conservative. As for the tests for $H_{01} : p_m = p_f$, LRT_1 and Z_1 almost have the same simulated power just like Fig 1. On the other hand, the powers of LRT_0, LRT_{0b}, Z_0, LRT_1 and Z_1 are not so affected by the ratio r. However, their powers are greatly influenced by the absolute difference $|\epsilon| = |p_m - p_f|$. Specifically, their powers under $p_f = 0.25$ and $p_f = 0.35$ are much larger than those under $p_f = 0.26$ and $p_f = 0.34$. In addition, when the simulation setting is fixed, LRT_1 and

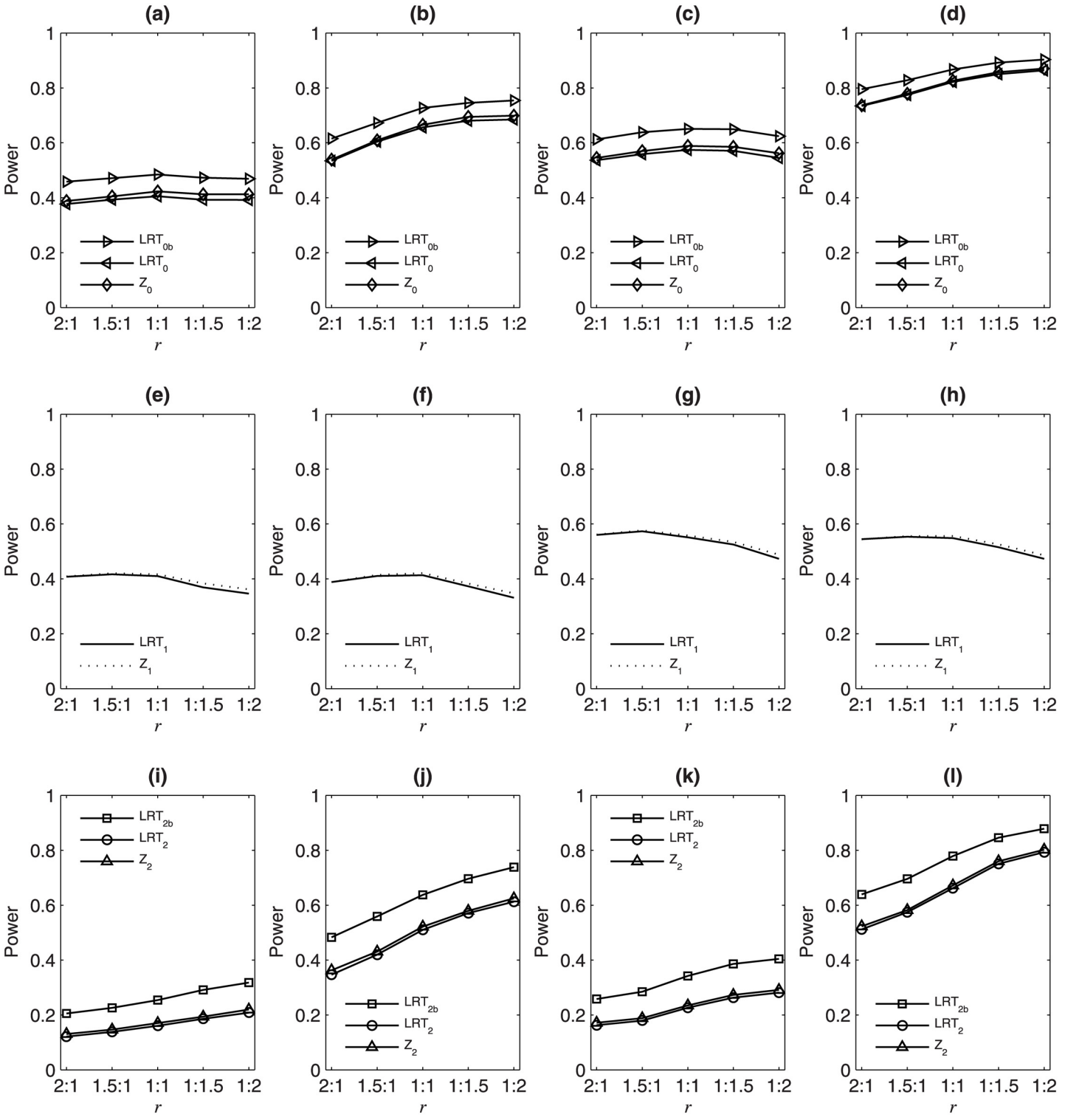

Fig 1. Simulated powers of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 against $r = N_m : N_f$ under $H_1 : p_m \neq p_f$ and $\rho > 0$ based on 10000 replicates with $p_m = 0.3$ and $p_f = 0.35$. In the first column: $\rho = 0.05$ and $N = 800$; in the second column: $\rho = 0.1$ and $N = 800$; in the third column: $\rho = 0.05$ and $N = 1200$; in the fourth column: $\rho = 0.1$ and $N = 1200$. In the first row, the powers of LRT_0, LRT_{0b} and Z_0 for $H_0 : p_m = p_f$ and $\rho = 0$; in the second row, the powers of LRT_1 and Z_1 for $H_{01} : p_m = p_f$; in the third row, the powers of LRT_2, LRT_{2b} and Z_2 for $H_{02} : \rho = 0$.

doi:10.1371/journal.pone.0145032.g001

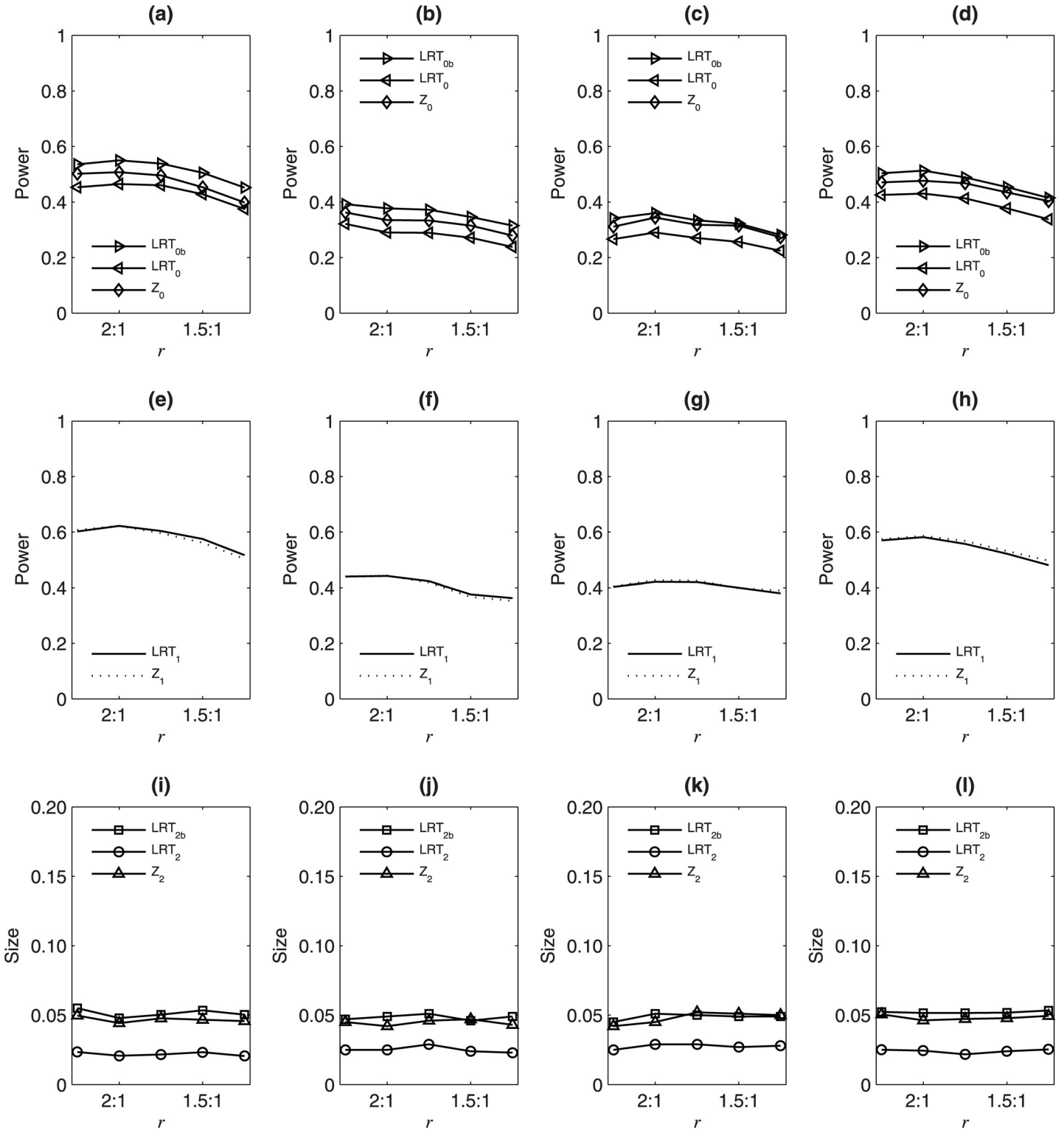

Fig 2. Simulated size/powers of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 against $r = N_m : N_f$ under $H_{02} : \rho = 0$ based on 10000 replicates with $p_m = 0.3$ and $N = 1200$. In the first column: $p_f = 0.25$; in the second column: $p_f = 0.26$; in the third column: $p_f = 0.34$; in the fourth column: $p_f = 0.35$. In the first row, the powers of LRT_0, LRT_{0b} and Z_0 for $H_0 : p_m = p_f$ and $\rho = 0$; in the second row, the powers of LRT_1 and Z_1 for $H_{01} : p_m = p_f$; in the third row, the size of LRT_2, LRT_{2b} and Z_2 for $H_{02} : \rho = 0$.

doi:10.1371/journal.pone.0145032.g002

Z_1 with one degree of freedom are a little more powerful than LRT_0, LRT_{0b} and Z_0 with two degrees of freedom, because the true model is $p_m \neq p_f$ and $\rho = 0$. By comparing Fig 2d ($\rho = 0$), Fig 1c ($\rho = 0.05$) and Fig 1d ($\rho = 0.1$) under $N = 1200$, $p_m = 0.3$ and $p_f = 0.35$, LRT_0, LRT_{0b} and Z_0 are more and more powerful with ρ increasing.

Figs A–G in S1 File show the corresponding results under other simulation settings with $p_m \neq p_f$ and $\rho > 0$, which are similar to those in Fig 1. Figs H and I in S1 File plot the corresponding results under $p_m = p_f$ and $\rho > 0$, and Figs J–L in S1 File give the corresponding results under $p_m \neq p_f$ and $\rho = 0$. The more details refer to S1 File.

Table 2 shows the simulated size of LRT_2, LRT_{2b} and Z_2 for $H_{02} : \rho = 0$ for different values of p_f, having $N_m = 0$ under the sample sizes N_f required for LRT_{2b} to obtain 80% simulated power. Table 3 lists the simulated powers under these sample sizes for different values of p_f, having $\rho =$ 0.05 and 0.1. From Table 2, we can see that the type I error rates of LRT_2, LRT_{2b} and Z_2 are close to the nominal significance level of 5%. It is shown in Table 3 that the power of LRT_{2b} attains to about 80%, and the difference in power between LRT_{2b} and Z_2 is about 10%.

Tables A–J in S1 File list the RMSEs and biases of the estimates of p_m, p_f, the pooled allele frequency p and ρ for different values of p_m, p_f, ρ, r and N. It should be noted that the estimate of p_m based on the EM algorithm is the same as zheng et al. [14]. Further, the estimates $\hat{p}_{f1}$ and $\hat{p}_{01}$ of p_f and p based on the EM algorithms have the similar RMSEs and biases as those from zheng et al. [14], respectively. However, when we focus on the estimate of ρ, we find that although the biases of $\hat{\rho}_1$ and $\hat{\rho}_{01}$ based on the EM algorithms are larger than $\hat{\rho}_z$ in zheng et al. [14] for some cases, the RMSEs of $\hat{\rho}_1$ and $\hat{\rho}_{01}$ are smaller than $\hat{\rho}_z$ for all the simulation settings.

Table 4 displays the simulated size/powers of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 under the population stratification model. When $\epsilon_1 = \epsilon_2 = 0$, the size of LRT_1 and Z_1 is obtained. Further, note that the ratios of two subpopulations in the whole population are equal. As such, $\epsilon_1 = -\epsilon_2$ will also cause the size of LRT_1 and Z_1. Under other simulation settings, we get the powers of the eight test statistics. To investigate whether or not the population stratification model

Table 2. Simulated size (in %) of LRT_2, LRT_{2b} and Z_2, having $N_m = 0$ and $\rho = 0$.

N_f	p_f	LRT_2	LRT_{2b}	Z_2
2500	0.20	2.27	5.18	4.84
	0.25	2.27	4.96	4.89
	0.30	2.43	5.15	5.13
	0.35	2.31	4.99	5.14
	0.40	2.25	4.96	4.70
	0.45	2.43	5.05	4.93
	0.50	2.54	5.03	5.06
	0.55	2.38	5.26	4.86
	0.60	2.49	5.11	5.10
650	0.20	2.14	4.98	4.67
	0.25	2.03	4.62	4.82
	0.30	2.53	4.91	5.20
	0.35	2.45	5.05	5.11
	0.40	2.18	4.72	4.54
	0.45	2.04	4.84	4.65
	0.50	2.48	4.98	5.14
	0.55	2.30	4.74	5.12
	0.60	2.27	5.01	4.77

doi:10.1371/journal.pone.0145032.t002

Table 3. Simulated powers (in %) of LRT_2, LRT_{2b} and Z_2, having $N_m = 0$.

N_f	ρ	p_f	LRT_2	LRT_{2b}	Z_2
2500	0.05	0.20	68.2	79.0	69.7
	0.05	0.25	69.1	79.4	69.9
	0.05	0.30	69.4	79.9	70.2
	0.05	0.35	69.8	80.2	70.3
	0.05	0.40	70.0	80.6	70.4
	0.05	0.45	69.3	79.9	69.6
	0.05	0.50	70.6	80.8	71.5
	0.05	0.55	71.2	81.0	71.5
	0.05	0.60	70.2	80.6	70.5
650	0.10	0.20	68.1	78.9	70.3
	0.10	0.25	69.2	79.9	70.8
	0.10	0.30	69.9	80.9	71.0
	0.10	0.35	70.3	80.5	71.1
	0.10	0.40	71.1	81.1	71.8
	0.10	0.45	71.9	81.9	72.5
	0.10	0.50	71.1	81.6	72.1
	0.10	0.55	70.7	81.4	71.3
	0.10	0.60	70.8	81.9	71.6

doi:10.1371/journal.pone.0145032.t003

causes excess homozygosity, we save the values of 10000 ρ estimates for each estimation method ($\hat{\rho}_1$, $\hat{\rho}_{01}$ or $\hat{\rho}_z$). Then, calculate the corresponding mean and standard deviation (SD), which are also listed in Table 4. The results show that the population stratification model indeed leads to the positive inbreeding coefficient (i.e., excess homozygosity), which is consistent with Overall and Nichols [24]. The mean $\hat{\rho}$ values ($\hat{\rho}_1$ and $\hat{\rho}_{01}$) using the EM algorithm are a little larger than $\hat{\rho}_z$ proposed in zheng et al. [14], while $\hat{\rho}_1$ and $\hat{\rho}_{01}$ have less standard deviation. On the other hand, the size of LRT_1 and Z_1 is close to the nominal significance level of 5%. The power of LRT_{0b} is larger than LRT_0 and Z_0, and LRT_0 and Z_0 have the similar powers, irrespective of the ϵ_1 and ϵ_2 values. LRT_1 and Z_1 have almost the same powers. LRT_{2b} is much more powerful than LRT_2 and Z_2, and the power of LRT_2 is a little smaller than Z_2. If ϵ_1 is fixed and ϵ_2 is changed, the ρ estimate increases with ϵ_2 increasing, and hence LRT_2, LRT_{2b} and Z_2 are more and more powerful; if ϵ_2 is fixed and ϵ_1 is changed, the ρ estimate decreases with ϵ_1 increasing, and hence LRT_2, LRT_{2b} and Z_2 are less and less powerful. This may be caused by p_m being taken to be 0.3 and 0.5 in the first and second subpopulations, respectively.

Application to RA data

We apply the proposed methods to the RA dataset from North American Rheumatoid Arthritis Consortium for studying their practicability, which is available from Genetic Analysis Workshop 15. In this dataset, there are 1217 families. Note that many individuals' genotypes are missing. On the other hand, to obtain a sample of which all the individuals are independent, we only select the available founders in this dataset, which results in a sample composed of 369 founders ($N_m = 112$ and $N_f = 257$) in the analysis. 293 SNP markers on X chromosome for each founder are included in this application. The significance level is fixed at $\alpha = 5\%$. Table 5 gives the corresponding results based on the P-values of LRT_{0b}, LRT_1, LRT_{2b}, Z_0, Z_1 and Z_2. From Table 5, LRT_{0b} identified 6 loci which Z_0 did not identify, and Z_0 identified 4 additional loci. One locus is detected by LRT_1 that is not found by Z_1, and 4 additional loci are detected by

Table 4. Mean and standard deviation (SD) of ρ estimates over 10000 replications, and simulated size/powers (in %) of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 under population stratification model.

ϵ		$\hat{\rho}_1$		$\hat{\rho}_{01}$		$\hat{\rho}_z$		Power							
ϵ_1[a]	ϵ_2[b]	Mean	SD	Mean	SD	Mean	SD	LRT_0	LRT_{0b}	LRT_1	LRT_2	LRT_{2b}	Z_0	Z_1	Z_2
-0.05	-0.05	0.043	0.031	0.046	0.032	0.042	0.034	72.3	78.6	72.1	24.1	37.0	72.6	71.5	25.0
	-0.04	0.050	0.031	0.052	0.032	0.049	0.034	67.8	74.4	63.4	31.0	43.9	67.7	63.0	31.6
	0.00	0.067	0.032	0.068	0.032	0.067	0.033	54.8	63.4	22.7	51.7	65.2	55.3	22.5	52.5
	0.04	0.088	0.034	0.088	0.034	0.087	0.034	65.0	72.3	4.7	73.6	82.8	65.6	4.7	74.5
	0.05	0.093	0.035	0.093	0.035	0.093	0.035	70.9	76.3	4.1	78.7	86.2	71.3	4.1	79.0
-0.04	-0.05	0.039	0.029	0.041	0.030	0.037	0.033	60.5	69.0	63.2	19.2	28.5	60.9	63.2	19.8
	-0.04	0.045	0.030	0.046	0.031	0.043	0.033	57.6	63.8	50.5	25.6	37.6	57.6	49.9	26.2
	0.00	0.061	0.033	0.062	0.033	0.060	0.034	42.2	51.1	17.0	44.5	57.7	42.6	17.0	45.0
	0.04	0.081	0.033	0.081	0.033	0.081	0.034	57.4	64.0	4.2	65.9	77.9	58.1	4.3	66.4
	0.05	0.088	0.033	0.088	0.033	0.088	0.033	65.2	72.5	5.0	74.7	85.2	65.8	5.0	75.7
0.00	-0.05	0.029	0.027	0.030	0.027	0.025	0.033	21.8	29.0	23.0	11.0	19.0	22.5	22.9	12.1
	-0.04	0.030	0.027	0.031	0.027	0.026	0.033	18.9	24.6	17.3	12.5	20.0	19.6	17.2	12.9
	0.00	0.043	0.029	0.043	0.029	0.041	0.032	17.7	24.0	4.3	23.1	34.9	18.2	4.3	23.9
	0.04	0.058	0.032	0.058	0.032	0.058	0.033	42.3	50.6	19.0	40.0	54.1	43.1	19.0	40.6
	0.05	0.063	0.032	0.063	0.032	0.063	0.033	49.4	57.4	23.3	46.2	59.5	50.1	23.5	46.9
0.04	-0.05	0.019	0.023	0.020	0.023	0.012	0.032	6.7	9.4	6.8	4.8	10.0	7.6	6.9	6.1
	-0.04	0.022	0.024	0.022	0.025	0.014	0.034	6.3	9.0	4.7	7.1	11.7	7.0	4.7	7.7
	0.00	0.031	0.028	0.031	0.028	0.026	0.034	18.0	23.7	17.5	11.8	20.7	18.9	17.5	12.8
	0.04	0.041	0.031	0.041	0.031	0.039	0.034	52.1	61.9	51.8	21.2	32.0	52.7	52.1	21.5
	0.05	0.046	0.031	0.047	0.031	0.045	0.034	61.7	69.5	58.6	27.1	38.9	62.1	58.8	27.6
0.05	-0.05	0.018	0.023	0.018	0.023	0.008	0.034	4.1	7.2	4.9	4.5	8.4	5.2	5.0	5.5
	-0.04	0.020	0.023	0.020	0.023	0.011	0.034	6.0	8.4	5.2	4.2	10.1	7.5	5.2	5.9
	0.00	0.028	0.028	0.028	0.028	0.022	0.035	22.1	28.2	24.1	10.4	16.9	22.9	24.2	11.2
	0.04	0.036	0.028	0.037	0.028	0.034	0.032	59.5	65.7	61.6	17.0	27.1	60.0	61.9	17.7
	0.05	0.042	0.031	0.043	0.030	0.040	0.033	68.1	74.1	67.2	22.6	32.1	68.3	67.6	23.4

[a] ϵ in the first subpopulation.

[b] ϵ in the second subpopulation.

doi:10.1371/journal.pone.0145032.t004

Z_1. There are 12 loci identified by LRT_{2b}, which can not be identified by Z_2, and only 2 additional loci are identified by Z_2. However, there exist multiple testing problems because we simultaneously analyze 293 loci. So, Bonferroni correction is used ($\alpha' = 0.05/293 = 1.71 \times 10^{-4}$) and there is no statistically significant result to occur. The more details can be found in Tables K–M in S1 File.

To investigate the computational efficiency of the XHWE software, we implement the code with the default arguments for this dataset (1217 families and 293 SNPs), on a HP 2311f personal computer (Microsoft Windows 7 Enterprise (Service Pack 1), 4GB of RAM and 3.40 GHz Intel(R) Core(TM) i7 Duo processor) and record its computational time. This process needs 977 seconds. Therefore, on the average, the running time for a single SNP is about 3.3 seconds. For the genome-wide case, for example, one would analyze 200000 SNP markers on X chromosome for the family sample of the type mentioned above, which would lead to 1600000 tests for the hypotheses with running time being about 7.6 days on the personal computer of this type.

Table 5. LRT_{0b}, LRT_1, LRT_{2b}, Z_0, Z_1 and Z_2 results of application to rheumatoid arthritis data at 5% level.

A. Contingency table showing LRT_{0b} and Z_0 results at 5% level.			
	$P_{Z_0} < 0.05$	$P_{Z_0} \geq 0.05$	**Total**
$P_{LRT_{0b}} < 0.05$	11	6	17
$P_{LRT_{0b}} \geq 0.05$	4	272	276
Total	15	278	293
B. Contingency table showing LRT_1 and Z_1 results at 5% level.			
	$P_{Z_1} < 0.05$	$P_{Z_1} \geq 0.05$	**Total**
$P_{LRT_1} < 0.05$	9	1	10
$P_{LRT_1} \geq 0.05$	4	279	283
Total	13	280	293
C. Contingency table showing LRT_{2b} and Z_2 results at 5% level.			
	$P_{Z_2} < 0.05$	$P_{Z_2} \geq 0.05$	**Total**
$P_{LRT_{2b}} < 0.05$	14	12	26
$P_{LRT_{2b}} \geq 0.05$	2	265	267
Total	16	277	293

doi:10.1371/journal.pone.0145032.t005

Discussion

The existing Z_1 and Z_2 tests were respectively proposed to test for $H_{01} : p_m = p_f$ and $H_{02} : \rho = 0$. However, we find that there is no simulation study conducted to assess the validity of Z_1 and Z_2 and their performance [14]. Further, there is no existing method to simultaneously test for $H_0 : p_m = p_f$ and $\rho = 0$. Therefore, in this article, we first combine these two test statistics and suggest $Z_0 = Z_1 + Z_2$ to test for the equality of the frequencies of the same allele in males and females and the zero inbreeding coefficient on X chromosome based on the collected sample, because Z_1 and Z_2 are independent of each other. What's more, for the purpose of improving the test power, we propose several LRT-type test statistics. Firstly, we write out the likelihood functions under $H_0 : p_m = p_f$ and $\rho = 0$ and $H_1 : p_m \neq p_f$ or $\rho > 0$ at a single SNP locus on X chromosome, respectively. Then, we obtain the MLEs of the male allele frequency, the female allele frequency and the inbreeding coefficient by the EM algorithms, where we use the RMSE and bias to assess the accuracy of the MLEs of these unknown parameters and construct the corresponding likelihood ratio test (LRT_0) statistic under the null hypothesis H_0. If H_0 is statistically rejected, we further develop two LRT-type test statistics LRT_1 and LRT_2 respectively for $H_{01} : p_m = p_f$ and $H_{02} : \rho = 0$. Note that LRT_0 and LRT_2 are too conservative from the simulated results. So, we use parametric bootstrap techniques and propose the LRT_{0b} and LRT_{2b} test statistics. We simulate the data under different parameter settings. Simulation results show that the proposed bootstrap-based methods LRT_{0b} and LRT_{2b}, LRT_1, Z_0 and the existing Z_1 and Z_2 control the type I error rates well under the respective null hypothesis. Power comparison demonstrates that LRT_{0b} is more powerful than both LRT_0 and Z_0. Under $\rho > 0$, LRT_{2b} has much more power than LRT_2 and Z_2, and LRT_2 is a little less powerful than Z_2. In addition, LRT_1 and Z_1 almost have the same power under $p_m \neq p_f$.

As for the parameter estimates, the estimate of p_m based on the EM algorithm is the same as that in zheng et al. [14]. Further, the estimates $\hat{p}_{f1}$ and $\hat{p}_{01}$ of p_f and the pooled allele frequency p based on the EM algorithms have the RMSE and bias similar to those from zheng et al. [14], respectively. However, although the biases of $\hat{\rho}_1$ and $\hat{\rho}_{01}$ based on the EM algorithms are larger than $\hat{\rho}_z$ from zheng et al. [14] for some cases, the RMSEs of $\hat{\rho}_1$ and $\hat{\rho}_{01}$ are smaller than $\hat{\rho}_z$ for all the simulation settings. In addition, the population stratification model indeed causes excess homozygosity, which is consistent with Overall and Nichols [24]. The mean $\hat{\rho}$ values ($\hat{\rho}_1$ and $\hat{\rho}_{01}$)

Table 6. Simulated size/powers (in %) of LRT_0, LRT_{0b}, LRT_1, LRT_{1b}, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 based on 10000 Monte Carlo replications and 1000 bootstrap replications under X chromosome inactivation and dose compensation, having $p_m = 0.3$ and the ratio $N_m : N_f = 1:1$.

N	ρ	p_f	LRT_0	LRT_{0b}	LRT_1	LRT_{1b}	LRT_2	LRT_{2b}	Z_0	Z_1	Z_2
800	0.00	0.30	6.5	5.0	10.6	4.9	2.3	5.2	4.8	4.8	4.9
	0.05	0.30	17.2	13.3	10.8	4.9	16.2	25.9	13.0	5.0	17.0
	0.10	0.30	44.1	38.5	11.4	5.2	49.3	63.0	41.2	5.3	50.5
	0.00	0.34	31.7	27.3	42.2	29.2	2.5	5.4	23.5	29.7	5.2
	0.05	0.34	41.8	36.5	41.2	27.9	15.8	26.3	31.6	28.4	16.4
	0.10	0.34	65.1	59.5	41.0	28.2	50.9	64.0	58.5	28.5	52.0
	0.00	0.35	43.6	38.7	55.5	41.3	2.2	4.9	33.6	41.9	5.0
	0.05	0.35	54.3	48.9	55.3	41.5	15.8	26.0	42.0	41.9	16.7
	0.10	0.35	72.1	67.3	53.7	40.2	49.9	63.1	64.8	40.6	50.9
1200	0.00	0.30	6.7	4.8	10.9	5.1	2.2	4.8	5.1	5.0	5.5
	0.05	0.30	22.4	18.0	10.8	5.1	23.0	34.3	18.4	5.0	24.3
	0.10	0.30	60.9	54.8	11.2	5.5	67.3	78.5	58.6	5.5	68.3
	0.00	0.34	42.7	37.8	55.0	40.7	2.2	4.9	32.5	41.3	4.9
	0.05	0.34	57.3	52.1	54.4	40.8	22.1	33.6	46.6	41.0	22.9
	0.10	0.34	81.0	76.8	53.0	39.3	67.5	79.3	76.0	39.5	68.7
	0.00	0.35	59.5	53.8	70.6	57.4	2.3	4.9	47.5	57.9	5.0
	0.05	0.35	71.2	66.2	70.3	57.5	22.7	34.3	59.7	57.8	23.8
	0.10	0.35	88.1	84.7	68.5	55.4	67.7	78.8	83.6	56.0	68.7

doi:10.1371/journal.pone.0145032.t006

using the EM algorithm are a little larger than $\hat{\rho}_z$ proposed in zheng et al. [14], while $\hat{\rho}_1$ and $\hat{\rho}_{01}$ have less standard deviation.

Note that $\rho = 0$ and $\rho > 0$ in the null and alternative hypotheses of the likelihood ratio test LRT_0 or LRT_2, respectively, which causes the "boundary" problem and that the corresponding likelihood ratio test is not expected to follow a χ^2 distribution [31, 33]. This may be the reason why the size of LRT_0 and LRT_2 is too conservative. Therefore, we use parametric bootstrap techniques to obtain the exact distributions of LRT_0 and LRT_2 in this article.

Due to the presence of the X chromosome inactivation (XCI) and dosage compensation (DC), association analysis and excess homozygosity tests on X chromosome are more complicated than those on autosomes [34]. In the presence of XCI, only one allele from a pair of alleles in females is expressed [35]. Consequently, if considering a locus with two alleles M_1 and M_2, the effect of the M_1 allele in males should be equivalent to the difference between M_2M_2 and M_1M_1 homozygous females. As such, when we conduct analyses based on allele-counting, we must either count each allele twice in males or equivalently count each allele in females as 0.5, reflecting a "dosage compensation" for X inactivation [34]. It should be noted that LRT_2, LRT_{2b} and Z_2 for $H_{02} : \rho = 0$ are not affected by XCI and DC because they only use female individuals in the collected sample. Similarly, Z_1 for $H_{01} : p_m = p_f$ is also not influenced by XCI and DC because it estimates the allele frequencies and the corresponding variances in males and females, respectively. Thus, $Z_0 = Z_1 + Z_2$ is still valid when XCI and DC exist. To investigate the effect of XCI and DC on LRT_0, LRT_{0b}, LRT_1 and LRT_{1b}, where LRT_{1b} is the bootstrap version of LRT_1, we carry out simulation study under several simulation settings in the presence of XCI and DC. The simulation settings and simulation results are listed in Table 6. It is shown in the table that the size of LRT_{2b}, Z_0, Z_1 and Z_2 stays close to the nominal 5% level and the size of LRT_2 is still conservative. However, LRT_0 and LRT_1 without bootstrap cannot control the size well. Fortunately, the type I error rates of LRT_{0b} and LRT_{1b} with bootstrap are very close to 5%. Furthermore, LRT_{0b} is more powerful than Z_0 almost for all the cases and LRT_{1b} and Z_1 almost

have the same performance in power. Therefore, in the presence of XCI and DC, LRT_{0b}, Z_1 and LRT_{2b} are recommended. Finally, LRT_{0b} and LRT_{2b} can deal with samples of small size. However, LRT_{0b} and LRT_{2b} are based on the parametric bootstrap techniques, which are more computationally intensive.

Supporting Information

S1 File. Supporting Information. Tables A–J, root mean squared errors (RMSE) and biases of estimates of p_m, p_f and ρ based on EM algorithm and zheng et al. [14] under different simulation settings. Tables K–M, LRT_0, LRT_{0b}, Z_0, LRT_1, Z_1, LRT_2, LRT_{2b}, and Z_2 results of application to rheumatoid arthritis data, respectively. Figs A–L, simulated size/powers of LRT_0, LRT_{0b}, LRT_1, LRT_2, LRT_{2b}, Z_0, Z_1 and Z_2 against $r = N_m : N_f$ based on 10000 replicates under different simulation settings.
(PDF)

Author Contributions

Conceived and designed the experiments: XPY JYZ. Performed the experiments: XPY JYZ. Analyzed the data: XPY QLZ JLL. Contributed reagents/materials/analysis tools: XPY QLZ JLL JYZ. Wrote the paper: XPY JYZ.

References

1. Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype—phenotype associations. Eur J Hum Genet. 2001; 9: 301–306. doi: 10.1038/sj.ejhg.5200625 PMID: 11313775
2. Lenart BA, Neviaser AS, Lyman S, Chang CC, Edobor-Osula F, Steele B, et al. Association of low-energy femoral fractures with prolonged bisphosphonate use: a case control study. Osteoporosis Int. 2009; 20: 1353–1362. doi: 10.1007/s00198-008-0805-x
3. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol. 2001; 20: 4–16. doi: 10.1002/1098-2272(200101)20:1%3C4::AID-GEPI2%3E3.0.CO;2-T PMID: 11119293
4. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genome wide association analysis of coronary artery disease. N Engl J Med. 2007; 357: 443–453. doi: 10.1056/NEJMoa072366 PMID: 17634449
5. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010; 42: 579–589. doi: 10.1038/ng.609 PMID: 20581827
6. Wang M, Lin S. FamLBL: detecting rare haplotype disease association based on common SNPs using case-parent triads. Bioinformatics. 2014; 30: 2611–2618. doi: 10.1093/bioinformatics/btu347 PMID: 24849576
7. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008; 40: 575–583. doi: 10.1038/ng.121 PMID: 18391952
8. Zhang H, Wheeler W, Wang Z, Taylor PR, Yu K. A fast and powerful tree-based association test for detecting complex joint effects in case-control studies. Bioinformatics. 2014; 30: 2171–2178. doi: 10.1093/bioinformatics/btu186 PMID: 24794927
9. Schaid DJ, Jacobsen SJ. Blased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. Am J Epidemiol. 1999; 149: 706–711. doi: 10.1093/oxfordjournals.aje.a009878 PMID: 10206619
10. Chung RH, Morris RW, Zhang L, Li YJ, and Martin RR. X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. Am J Hum Genet. 2007; 80: 59–68. doi: 10.1086/510630 PMID: 17160894
11. Ding J, Lin S, Liu Y. Monte Carlo pedigree disequilibrium test for markers on the X chromosome. Am J Hum Genet. 2006; 79: 567–573. doi: 10.1086/507609 PMID: 16909396

12. Horvath S, Laird NM, Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. Am J Hum Genet. 2000; 66: 1161–1167. doi: 10.1086/302823 PMID: 10712229
13. Zhang L, Martin ER, Chung RH, Li YJ, Morris RW. X-LRT: a likelihood approach to estimate genetic risks and test association with X-linked markers using a case-parents design. Genet Epidemiol. 2008; 32: 370–380. doi: 10.1002/gepi.20311 PMID: 18278816
14. Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. Genet Epidemiol. 2007; 31: 834–843. doi: 10.1002/gepi.20244 PMID: 17549761
15. Clayton D. Testing for association on the X chromosome. Biostat. 2008; 9: 593–600. doi: 10.1093/biostatistics/kxn007
16. Hickey PF, Bahlo M. X chromosome association testing in genome wide association studies. Genet Epidemiol. 2011; 35: 664–670. doi: 10.1002/gepi.20616 PMID: 21818774
17. 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. 2014; doi: 10.1177/0962280214551815
18. Li CC. First Course in Population Genetics. Boxwood Press; 1976.
19. Gillespie JH. Population Genetics: A Concise Guide. 1st ed. Johns Hopkins University Press; 2010.
20. König IR, Loley C, Erdmann J, Ziegler A. How to include chromosome X in your genome-wide association study. Genet Epidemiol. 2014; 38: 97–103. doi: 10.1002/gepi.21782
21. Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PLoS ONE. 2014; 9: e113684. doi: 10.1371/journal.pone.0113684 PMID: 25479423
22. Wright S. Systems of mating. Genetics. 1921; 6: 111–178. PMID: 17245958
23. Nei M. Molecular evolutionary genetics. Columbia University Press, New York; 1987.
24. Overall ADJ, Nichols RA. A method for distinguishing consanguinity and population substructure using multilocus genotype data. Mol Biol Evol. 2001; 18: 2048–2056. doi: 10.1093/oxfordjournals.molbev.a003746 PMID: 11606701
25. Brookfield JFY. A simple new method for estimating null allele frequency from heterozygote deficiency. Mol Ecol. 1996; 5: 453–455. PMID: 8688964
26. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977; 39: 1–38.
27. Mao WG, He HQ, Xu Y, Chen PY, Zhou JY. Powerful haplotype-based Hardy-Weinberg equilibrium tests for tightly linked loci. PLoS ONE. 2013; 8: e77399. doi: 10.1371/journal.pone.0077399 PMID: 24167573
28. Emigh TH. A comparison of tests for Hardy-Weinberg equilibrium. Biometrics. 1980; 36: 627–642. doi: 10.2307/2556115 PMID: 25856832
29. Kuk AYC, Zhang H, Yang Y. Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium. Bioinformatics. 2009; 25: 379–386. doi: 10.1093/bioinformatics/btn623 PMID: 19050036
30. Gartler SM. A brief history of dosage compensation. J Genet. 2014; 93: 591–595. doi: 10.1007/s12041-014-0360-5 PMID: 25189265
31. Weir BS, Cockerham CC. Complete Characterization of Disequilibrium at Two Loci. Mathematical Evolutionary Theory, Princeton University Press; 1989.
32. Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 1987; 82: 605–610. doi: 10.1080/01621459.1987.10478472
33. Zhang DW, Lin XH. Hypothesis testing in semiparametric additive mixed models. Biostat. 2003; 4: 57–74. doi: 10.1093/biostatistics/4.1.57
34. Bielawski JP, Yang Z. Maximum Likelihood Methods for Detecting Adaptive Protein Evolution. Statistical Methods in Molecular Evolution. Springer; 2005.
35. Clayton DG. Sex chromosomes and genetic association studies. Genome Med. 2009; 1: 110. doi: 10.1186/gm110 PMID: 19939292
36. Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, et al. X chromosome-inactivation patterns in 1,005 phenotypically unaffected females. Am J Hum Genet. 2006; 79: 493–499. doi: 10.1086/507565 PMID: 16909387