Powerful Haplotype-Based Hardy-Weinberg Equilibrium Tests for Tightly Linked Loci

Wei-Gao Mao, Hai-Qiang He, Yan Xu, Ping-Yan Chen, Ji-Yuan Zhou*

Department of Biostatistics, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong, China

Abstract

Recently, there have been many case-control studies proposed to test for association between haplotypes and disease, which require the Hardy-Weinberg equilibrium (HWE) assumption of haplotype frequencies. As such, haplotype inference of unphased genotypes and development of haplotype-based HWE tests are crucial prior to fine mapping. The goodness-of-fit test is a frequently-used method to test for HWE for multiple tightly-linked loci. However, its degrees of freedom dramatically increase with the increase of the number of loci, which may lack the test power. Therefore, in this paper, to improve the test power for haplotype-based HWE, we first write out two likelihood functions of the observed data based on the Niu's model (NM) and inbreeding model (IM), respectively, which can cause the departure from HWE. Then, we use two expectation-maximization algorithms and one expectation-conditional-maximization algorithm to estimate the model parameters under the HWE, IM and NM models, respectively. Finally, we propose the likelihood ratio tests LRT_1 and LRT_2 for haplotype-based HWE under the NM and IM models, respectively. We simulate the HWE, Niu's, inbreeding and population stratification models to assess the validity and compare the performance of these two LRT tests. The simulation results show that both of the tests control the type I error rates well in testing for haplotype-based HWE. If the NM model is true, then LRT_1 is more powerful. While, if the true model is the IM model, then LRT_2 has better performance in power. Under the population stratification model, LRT_2 is still more powerful. To this end, LRT_2 is generally recommended. Application of the proposed methods to a rheumatoid arthritis data set further illustrates their utility for real data analysis.

Citation: Mao W-G, He H-Q, Xu Y, Chen P-Y, Zhou J-Y (2013) Powerful Haplotype-Based Hardy-Weinberg Equilibrium Tests for Tightly Linked Loci. PLoS ONE 8(10): e77399. doi:10.1371/journal.pone.0077399

Editor: Claire Wade, University of Sydney, Australia

Received April 24, 2013; **Accepted** September 2, 2013; **Published** October 22, 2013

Funding: This work was supported by the National Natural Science Foundation of China (81072386) and grants from School of Public Health and Tropical Medicine of Southern Medical University, China (grant numbers: GW201219 and GW201237). 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-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.

* E-mail: zhoujiyuan@gmail.com

Introduction

In studies of genetic epidemiology, complex diseases are often associated with multiple (interacting) markers [1–3]. As such, haplotype-based analysis has gained increasing attention as it can potentially be more efficient than a single-marker-based analysis [4–9]. Therefore, haplotype inference of unphased genotypes may be expected to play an important role in disease fine mapping [10]. Nowadays, there are many statistical and computational methods available for inferring haplotypes based on different types of data, such as unrelated individuals. One of the popular approaches is the likelihood method, and the maximum likelihood estimation via the expectation-maximization (EM) algorithm [11] is a frequently employed method for haplotype inference. For genotype data of unrelated individuals, an EM-based maximum likelihood method for the estimation of haplotype frequencies was first proposed by Excoffier and Slatkin [12]. We call it EM algorithm in this paper for easy description later. However, the EM algorithm needs the assumption that the population under study is in Hardy-Weinberg equilibrium (HWE), otherwise the estimates of haplotype frequencies may be biased.

Recently, there have been many case-control studies proposed to test for association between haplotypes and disease. The likelihood ratio test (LRT) was constructed from the maximum likelihood functions for cases, controls and the pooled data of cases and controls, to test for haplotype-disease association, which requires the assumption of HWE in the pooled sample data [3]. Prospective likelihood methods based on logistic regression or generalized linear models were investigated by Schaid et al. [13], Stram et al. [14], Zaykin et al. [15], and others. These methods treat unobserved haplotypes as covariates in a regression model and compute the conditional expectation of the covariates given genotype observations under the null hypothesis of no association with a HWE assumption in the pooled sample of cases and controls. Zhao et al. [16] proposed a prospective estimating-equation approach for the assessment of disease association with haplotypes when adjustment for covariates, which needs the HWE assumption of haplotype frequencies only in the control sample. The pooled sample of cases and controls is not necessarily in HWE. On the other hand, a retrospective likelihood method can be used in detecting haplotype-disease association in a case-control study and also requires HWE only in the control population [17]. Therefore, the detection of haplotype-based HWE is crucial prior to fine mapping and positional cloning studies for case-control designs.

The goodness-of-fit test is a frequently-used method to test for HWE for multiple tightly-linked loci. However, when the number of loci under study increases, the degrees of freedom dramatically increase, which may lack the test power. As such, in this paper, to investigate more powerful haplotype-based HWE tests, we first recall three models which can cause Hardy-Weinberg disequilibrium (HWD). One was proposed originally by Niu et al. [6], which includes a parameter K and is called Niu's model (NM) in this paper for convenience; the second one is the inbreeding model (IM) with incorporating the inbreeding coefficient ρ [18]; the third one is a population stratification (PS) model, which can also lead to HWD. Then, we write out two likelihood functions of the observed data based on the NM and IM models, respectively. We develop an expectation-conditional-maximization (ECM) algorithm [19] for the NM model to estimate the parameter K and haplotype frequencies and suggest an EM algorithm for the IM model (denoted by IEM algorithm here) to estimate the inbreeding coefficient ρ and haplotype frequencies. Note that $K=1$ or $\rho=0$ means that HWE holds. So, we further propose two LRT tests LRT_1 and LRT_2 to test for haplotype-based HWE under the NM and IM models, respectively. We simulate the HWE, Niu's, inbreeding and population stratification models to assess the validity and compare the performance of these two LRT tests. The simulation results show that both of the tests control the size well in testing for haplotype-based HWE. If the Niu's model is true, then LRT_1 is more powerful. While, if the inbreeding model is true, then LRT_2 has better performance in power. Under the population stratification model, LRT_2 is still more powerful. Therefore, LRT_2 is generally recommended. In addition, we obtain the sum of absolute differences (SAD) between the true and estimated haplotype frequencies [20], and compare the performance of the EM, ECM and IEM algorithms in estimating the haplotype frequencies. If the true model is the Niu's model, then the ECM algorithm has more accurate estimates of haplotype frequencies than the EM and IEM estimates. However, for all the other simulation settings, the EM algorithm is not so much affected by the departure from HWE, and the EM and IEM algorithms almost have the same performance in controlling SAD, which is less than the ECM estimates. Application of the proposed methods to the Rheumatoid Arthritis (RA) data set from the North American Rheumatoid Arthritis Consortium (NARAC) further illustrates their utility for real data analysis.

Materials and Methods

Likelihood Function and EM Algorithm under HWE

Consider a sample of n unrelated individuals and q single nucleotide polymorphism (SNP) markers. Assume that the SNPs are tightly linked so that the recombination fraction between any SNP pair is zero. For each SNP, there are two alleles 1 and 2. Let $\mathbf{H}=\{h_1,h_2,\ldots,h_m\}$ be the set of all possible haplotypes at these q loci, where $m=2^q$. We assume that θ_i is the frequency of haplotype h_i $(i=1,2,\ldots,m)$, so the set of haplotype frequencies can be denoted by $\mathbf{\Theta}=\{\theta_1,\theta_2,\ldots,\theta_m\}$. Let $\mathbf{G}=\{G_1,G_2,\ldots,G_n\}$ be the set of the observed genotypes of all the n individuals, where G_j is the genotype of the j^{th} individual. For the j^{th} individual, the number of haplotype combinations compatible with G_j is s_j. Therefore, the likelihood function of the sample can be expressed as

$$L(\mathbf{\Theta})=L(\mathbf{G}|\mathbf{\Theta})=\prod_{j=1}^{n}P(G_j|\mathbf{\Theta})=\prod_{j=1}^{n}\sum_{k=1}^{s_j}P(H_{jk}|\mathbf{\Theta}) \qquad (1)$$

where H_{jk} denotes the k^{th} haplotype combination compatible with genotype G_j for the j^{th} individual.

To make the haplotype frequency estimation easy and feasible, the EM algorithm was employed [11]. Let $\mathbf{Z}=(Z_1,Z_2,\ldots,Z_n)$ be the true haplotype combinations of the sample which are actually unobserved, and Z_j is the true haplotype combination of the j^{th} individual. Then the log-likelihood function of the complete data is

$$l_c(\mathbf{\Theta})=\sum_{j=1}^{n}\sum_{k=1}^{s_j}I(Z_j=H_{jk})ln\left[P(H_{jk}|\mathbf{\Theta})\right] \qquad (2)$$

where $I(\cdot)$ is an indicator function and $I(\cdot)=1$ if $Z_j=H_{jk}$ and 0 otherwise. Note that under HWE, the probability $P(H_{jk}=h_rh_s|\mathbf{\Theta})$ of unordered haplotype pair h_rh_s is θ_r^2 if $r=s$ and $2\theta_r\theta_s$ otherwise. Further, Excoffier and Slatkin [12] proposed the following EM algorithm to obtain the maximum likelihood estimates of θ_i $(i=1,2,\ldots,m)$ at iteration $(t+1)$,

$$\hat{\theta}_i^{(t+1)}=\frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{s_j}\tau_{jk}^{i}\frac{P(H_{jk}|\hat{\mathbf{\Theta}}^{(t)})}{\sum_{l=1}^{s_j}P(H_{jl}|\hat{\mathbf{\Theta}}^{(t)})}, \quad t=0,1,2,\ldots$$

where τ_{jk}^{i} is the number of times that haplotype h_i occurs in the k^{th} haplotype combination for the j^{th} individual and takes values of 0, 1 or 2, and $P(H_{jk}|\hat{\mathbf{\Theta}}^{(t)})$ is the value of the probability $P(H_{jk}|\mathbf{\Theta})$ based on the estimated haplotype frequencies $\hat{\mathbf{\Theta}}^{(t)}=\{\hat{\theta}_1^{(t)},\hat{\theta}_2^{(t)},\ldots,\hat{\theta}_n^{(t)}\}$ at iteration t.

Two Forms of HWD

Note that the underlying assumption of HWE is strong and HWE does not hold usually. One may consider the following form of HWD,

$$P(H_{jk}=h_rh_s|\mathbf{\Theta})=\begin{cases}\theta_r^2+\rho\theta_r(1-\theta_r), & r=s\\ 2(1-\rho)\theta_r\theta_s, & r<s\end{cases} \qquad (3)$$

where ρ is the inbreeding coefficient which is generally positive [21]. Note that Equation (3) is reduced to HWE when $\rho=0$. We denote this form of HWD as "inbreeding model (IM)" for convenient description in this paper.

Another form of the departure from HWE was originally proposed by Niu et al. [6] as follows. Assume that the probability of unordered haplotype pair h_rh_s is proportional to $a\theta_r^2$ if $r=s$ and $2b\theta_r\theta_s$ otherwise, with two parameters a and b. Obviously, the HWE assumption holds if $a=b$. Note that the sum of all these terms for all the m haplotypes at the q loci may not be 1. Then, HWD can be defined as the following form:

$$P(H_{jk}=h_rh_s|\mathbf{\Theta})=\begin{cases}a\theta_r^2/T, & r=s\\ 2b\theta_r\theta_s/T, & r<s\end{cases} \qquad (4)$$

where

$$T=a\sum_{r=1}^{m}\theta_r^2+2b\sum_{r=1}^{m}\sum_{r<s}\theta_r\theta_s$$

Let $K=a/b$. Then, we assume $K\geq 1$ due to the positive inbreeding coefficient ρ. We denote this form of HWD as "Niu's model (NM)" for convenience.

Likelihood Function and Haplotype-Based HWE Test under Niu's Model

Using Equations (2) and (4), the log-likelihood function of the complete data under the Niu's model can be expressed as

$$l_{c1}(\mathbf{\Psi})=\sum_{j=1}^{n}\sum_{r=1}^{m}\sum_{r\leq s}I(Z_j=h_rh_s)ln\left[a\theta_r^2I(r=s)+2b\theta_r\theta_sI(r<s)\right] \\ -nlnT \tag{5}$$

where $\mathbf{\Psi}=(K,\mathbf{\Theta})$. In fact, there is only one additional parameter K included in Equation (5), compared to the likelihood function under HWE. So, we propose the following expectation-conditional-maximization (ECM) algorithm to estimate the haplotype frequencies and the parameter K. It consists of one expectation step (E-step) and m conditional-maximization steps (CM-steps) at each iteration. In E-step at iteration $(t+1)$, we can get the following Q function after taking the conditional expectation of Equation (5), given the observed genotype data $\mathbf{G}$ and current estimate $\hat{\mathbf{\Psi}}^{(t)}$ of $\mathbf{\Psi}$,

$$Q_1(\mathbf{\Psi};\hat{\mathbf{\Psi}}^{(t)})=\sum_{j=1}^{n}\sum_{r=1}^{m}\sum_{r\leq s}P(Z_j=h_rh_s|G_j,\hat{\mathbf{\Psi}}^{(t)}) \\ ln[a\theta_r^2I(r=s)+2b\theta_r\theta_sI(r<s)]-nln(T) \tag{6}$$

where $P(Z_j=h_rh_s|G_j,\mathbf{\Psi}^{(t)})$ is the conditional probability of the haplotype pair h_rh_s given G_j and $\hat{\mathbf{\Psi}}^{(t)}$, which is 0 if there is no haplotype pair compatible with genotype G_j.

In CM-steps, we maximize the Q function in Equation (6) to estimate $\mathbf{\Psi}$. Let $\hat{\mathbf{\Psi}}^{(x/m+t)}$ be the estimate of $\mathbf{\Psi}$ in the x^{th} CM-step among m CM-steps at iteration $(t+1)$. The detailed CM-steps are as follows:

- Give the initial value $\mathbf{\Psi}^{(0)}=(K^{(0)},\mathbf{\Theta}^{(0)})$, where $\mathbf{\Theta}^{(0)}=(\theta_1^{(0)},\theta_2^{(0)},\ldots,\theta_m^{(0)})$.
- At iteration $(t+1)$, by fixing $\hat{\mathbf{\Psi}}^{(t)}$ in the first CM-step, maximize the Q function by taking the first-order derivation with respect to K so as to get the estimate of K, and then

$$\hat{K}^{(t+1)}=\frac{2B_1^{(t)}\sum\limits_{r=1}^{m}\sum\limits_{r<s}\hat{\theta}_r^{(t)}\hat{\theta}_s^{(t)}}{B_2^{(t)}\sum\limits_{r=1}^{m}\left[\hat{\theta}_r^{(t)}\right]^2}$$

where $B_1^{(t)}=\sum_{j=1}^{n}\sum_{r=1}^{m}P(Z_j=h_rh_r|G_j,\hat{\mathbf{\Psi}}^{(t)})$, $B_2^{(t)}=\sum_{j=1}^{n}\sum_{r=1}^{m}\sum_{r<s}P(Z_j=h_rh_s|G_j,\hat{\mathbf{\Psi}}^{(t)})$. So, $\hat{\mathbf{\Psi}}^{(1/m+t)}=(\hat{K}^{(t+1)},\hat{\mathbf{\Theta}}^{(t)})$.

- Note that there is a constraint condition $\theta_1+\theta_2+\ldots+\theta_m=1$ when we maximize $Q_1(\hat{\mathbf{\Psi}}^{(1/m+t)};\hat{\mathbf{\Psi}}^{(t)})$ to estimate the haplotype frequencies $\mathbf{\Theta}$. Thus, from the second CM-step to the m^{th} CM-step, θ_i's $(i=2,3,\ldots,m)$ are estimated step by step and θ_1 is then estimated by $1-\hat{\theta}_2^{(t+1)}-\hat{\theta}_3^{(t+1)}-\ldots-\hat{\theta}_m^{(t+1)}$. Let $\hat{\mathbf{\Theta}}_{-x}^{(x/m+t)}$ be the set of the haplotype frequency estimates for all the haplotypes but h_1 and h_x in the x^{th} CM-step. Then, $\hat{\mathbf{\Theta}}_{-x}^{(x/m+t)}=\left(\hat{\theta}_2^{(t+1)},\hat{\theta}_3^{(t+1)},\ldots,\hat{\theta}_{(x-1)}^{(t+1)},\hat{\theta}_{x+1}^{(t)},\ldots,\hat{\theta}_m^{(t)}\right)$. For example, $\hat{\mathbf{\Theta}}_{-2}^{(2/m+t)}=(\hat{\theta}_3^{(t)},\hat{\theta}_4^{(t)},\ldots,\hat{\theta}_m^{(t)})$ in the second CM-step for estimating θ_2. As such, in the x^{th} CM-step $(x=2,3,\ldots,m)$, by maximizing $Q_1(\hat{\mathbf{\Psi}}^{((x-1)/m+t)};\hat{\mathbf{\Psi}}^{(t)})$, it is shown in Text S1 that a cubic equation with respect to $\hat{\theta}_x^{(t+1)}$ is obtained,

$$A\left[\hat{\theta}_x^{(t+1)}\right]^3+B\left[\hat{\theta}_x^{(t+1)}\right]^2+C\hat{\theta}_x^{(t+1)}+D=0 \tag{7}$$

where the coefficients A, B, C and D are, respectively,

$$A=2[\hat{K}^{(t+1)}-1](2n-C_x-C_1),$$
$$B=2[\hat{K}^{(t+1)}-1](1-A_1)(2C_x+C_1-3n)$$

$$C=2[\hat{K}^{(t+1)}-1](1-A_1)^2(n-C_x)-A_2(C_x+C_1),$$
$$D=(1-A_1)A_2C_x$$

$$C_y=\sum_{j=1}^{n}\sum_{r=1}^{m}2^{I(r=y)}P(Z_j=h_yh_r|G_j,\hat{\mathbf{\Psi}}^{(t)}),\ \ A_1=\left(\hat{\mathbf{\Theta}}_{-x}^{(x/m+t)}\right)\mathbf{E}$$

$$A_2=(\hat{K}^{(t+1)}-1)\left(\hat{\mathbf{\Theta}}_{-x}^{(x/m+t)}\right)\mathbf{F}\left(\hat{\mathbf{\Theta}}_{-x}^{(x/m+t)}\right)^T- \\ 2(\hat{K}^{(t+1)}-1)A_1+\hat{K}^{(t+1)}$$

and the vector $\mathbf{E}$ and the matrix $\mathbf{F}$ are respectively

$$\mathbf{E}=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix},\quad \mathbf{F}=\begin{pmatrix}2&1&1&\ldots&1\\1&2&1&\ldots&1\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&1&1&\ldots&2\end{pmatrix}$$

Moreover, the cubic equation above is alway solvable, and its solution can be obtained by Shengjin's formulas [22]. Note that the likelihood function converges no matter which initial values of $\mathbf{\Psi}$ are chosen. So, if there are two or three solutions between 0 and 1, then we can choose the solution which is closer to $\hat{\theta}_x^{(t)}$ in the former step. After this step,

$$\hat{\mathbf{\Psi}}^{(x/m+t)}=(\hat{K}^{(t+1)},\hat{\theta}_1^{(t)},\hat{\theta}_2^{(t+1)},\ldots,\hat{\theta}_x^{(t+1)},\hat{\theta}_{x+1}^{(t)},\ldots,\hat{\theta}_{m-1}^{(t)},\hat{\theta}_m^{(t)})$$

- For θ_1, $\hat{\theta}_1^{(t+1)}=1-\hat{\theta}_2^{(t+1)}-\hat{\theta}_3^{(t+1)}-\ldots-\hat{\theta}_m^{(t+1)}$. Then $\hat{\mathbf{\Psi}}^{(t+1)}=(\hat{K}^{(t+1)},\hat{\mathbf{\Theta}}^{(t+1)})$.
- Repeat the steps above until the observed log-likelihood function of Equation (1) converges.

Equation (1) can be written to be $L(\mathbf{\Psi})=\Pi_{j=1}^{n}\left[\sum_{k=1}^{s_j}P(H_{jk}|\mathbf{\Psi})\right]$ under the Niu's Model. Note that HWE holds when $K=1$ and HWE is violated otherwise. Therefore, a likelihood ratio test (LRT) for HWE is naturally constructed based on the estimated haplotype frequencies as follows,

$$LRT_1=2ln\left[\frac{L(\hat{\mathbf{\Psi}})}{L(\hat{\mathbf{\Psi}}_0)}\right]=2[lnL(\hat{\mathbf{\Psi}})-lnL(\hat{\mathbf{\Psi}}_0)] \tag{8}$$

where $L(\hat{\mathbf{\Psi}}_0)$ and $L(\hat{\mathbf{\Psi}})$ are the values of the observed likelihood function under the null hypothesis of HWE and under the HWD alternative, respectively. Obviously, this LRT statistic asymptotically follows a Chi-square distribution with the degree of freedom being 1 when HWE holds.

Likelihood Function and Haplotype-Based HWE Test under Inbreeding Model

Borrowing the idea of Zeng and Lin on how to estimate the haplotype frequencies based on case-control data for testing for association [18], here we rewrite the likelihood function for unrelated individuals under study and then propose a haplotype-based HWE test under the inbreeding model. Let H_j be a random variable, which takes values from s_j possible haplotype combinations compatible with G_j of the j^{th} individual. Suppose that $\rho \geq 0$, and W_j is a Bernoulli variable with success probability ρ. Let $P(V_{1j}=h_r/h_r)=\theta_r$ and $P(V_{2j}=h_r/h_s)=\theta_r\theta_s$, where V_{1j} and V_{2j} are discrete random variables, and the haplotype before "/" is paternal and haplotype after "/" is maternal. So, $W_jV_{1j}+(1-W_j)V_{2j}$ has the same distribution as H_j, and we treat W_j, V_{1j} and V_{2j} as missing. Then, the log-likelihood function of the complete data under the inbreeding model is

$$l_{c2}(\mathbf{\Phi})=\Pi_{j=1}^{n}\Big[\rho^{W_j}(1-\rho)^{1-W_j}\times \Pi_{r=1}^{m}\theta_r^{W_jI(V_{1j}=h_r/h_r)}\Pi_{r,s=1}^{m}(\theta_r\theta_s)^{(1-W_j)I(V_{2j}=h_r/h_s)}\Big] \tag{9}$$

where $\mathbf{\Phi}=(\rho,\mathbf{\Theta})$.

To estimate the parameters $\mathbf{\Psi}$ in Equation (9), the EM algorithm is considered. In E-step, the Q function is

$$Q_2(\mathbf{\Phi};\hat{\mathbf{\Phi}}^{(t)})=\sum_{j=1}^{n}\Big\{E(W_j|G_j,\hat{\mathbf{\Phi}}^{(t)})ln\rho+E(1-W_j|G_j,\hat{\mathbf{\Phi}}^{(t)})ln(1-\rho)$$

$$+\sum_{r=1}^{m}E[W_jI(V_{1j}=h_r/h_r)|G_j,\hat{\mathbf{\Phi}}^{(t)}]ln\theta_r$$

$$+\sum_{r,s=1}^{m}E[(1-W_j)I(V_{2j}=h_r/h_s)|G_j,\hat{\mathbf{\Phi}}^{(t)}]ln(\theta_r\theta_s)\Big\}$$

In M-step, the estimation of $\mathbf{\Phi}$ at iteration $(t+1)$ can be obtained by solving the following equation

$$\frac{\partial Q_2(\mathbf{\Phi};\hat{\mathbf{\Phi}}^{(t)})}{\partial \mathbf{\Phi}}=\mathbf{0}$$

So, ρ can be estimated by

$$\hat{\rho}^{(t+1)}=n^{-1}\sum_{j=1}^{n}E[W_j|G_j,\hat{\mathbf{\Phi}}^{(t)}]$$

$$=n^{-1}\sum_{j=1}^{n}\frac{\sum_{i=1}^{m}\hat{\rho}^{(t)}\hat{\theta}_i^{(t)}I(V_{1j}=h_i/h_i)}{\sum_{k=1}^{s_j}P(H_{jk}|G_j,\hat{\mathbf{\Phi}}^{(t)})}$$

Table 1. Haplotype distribution for Niu's model and inbreeding model.

SNP	Frequency
122	0.082
221	0.525
121	0.283
211	0.004
111	0.106

doi:10.1371/journal.pone.0077399.t001

where $\hat{\rho}^{(t)}$ and $\hat{\theta}_i^{(t)}$ are the estimates of ρ and θ_i at iteration t, respectively. The haplotype frequencies can be estimated by

$$\hat{\theta}_i^{(t+1)}=c^{-1}\sum_{j=1}^{n}\Big\{E[W_jI(V_{1j}=h_i/h_i|G_j,\hat{\mathbf{\Phi}}^{(t)})]+$$

$$2\sum_{r=1}^{m}E[(1-W_j)I(V_{2j}=h_i/h_r)|G_j,\hat{\mathbf{\Phi}}^{(t)}]\Big\}$$

where c is a normalizing constant, and $E[W_jI(V_{1j}=h_i/h_i)|G_j,\hat{\mathbf{\Phi}}^{(t)}]$ and $E[(1-W_j)I(V_{2j}=h_i/h_r)|G_j,\hat{\mathbf{\Phi}}^{(t)}]$ can be calculated as follows,

$$E[W_jI(V_{1j}=h_i/h_i)|G_j,\hat{\mathbf{\Phi}}^{(t)}]=\frac{\sum_{r=1}^{m}\hat{\rho}^{(t)}\hat{\theta}_r^{(t)}I(V_{1j}=h_i/h_i)}{\sum_{k=1}^{s_j}P(H_{jk}|G_j,\hat{\mathbf{F}}^{(t)})}$$

$$E[(1-W_j)I(V_{2j}=h_i/h_r)|G_j,\hat{\mathbf{\Phi}}^{(t)}]=$$

$$\frac{\sum_{f=1}^{m}\sum_{g=1}^{m}(1-\hat{\rho}^{(t)})\hat{\theta}_f^{(t)}\hat{\theta}_g^{(t)}I(V_{2j}=h_i/h_r)}{\sum_{k=1}^{s_j}P(H_{jk}|G_j,\hat{\mathbf{\Phi}}^{(t)})}$$

We call this process IEM algorithm for distinguishing it from the previous EM algorithm under HWE.

Table 2. Haplotype distribution for population stratification model.

SNP	Frequency	
	I	II
122	0.082	0.030
212	0.000	0.170
112	0.000	0.050
221	0.525	0.470
121	0.283	0.100
211	0.004	0.150
111	0.106	0.030

doi:10.1371/journal.pone.0077399.t002

Table 3. Mean and standard deviation (SD) of K and ρ estimates, mean of sum of absolute differences (SAD) of haplotype frequency estimates for EM, ECM and IEM algorithms, simulated size and powers of two HWE tests for different values of K and n, under Niu's model.

n	K	$\hat{K}$		$\hat{\rho}$		**SAD**			**Size/Power**	
		Mean	**SD**	**Mean**	**SD**	**EM**	**ECM**	**IEM**	**LRT$_1$**	**LRT$_2$**
500	1.00	1.001	0.107	0.009	0.016	0.041	0.043	0.041	0.054	0.029
	1.05	1.055	0.115	0.016	0.020	0.041	0.043	0.041	0.072	0.066
	1.10	1.099	0.118	0.022	0.022	0.043	0.043	0.043	0.135	0.107
	1.15	1.148	0.123	0.030	0.024	0.048	0.045	0.048	0.240	0.174
	1.20	1.198	0.124	0.039	0.027	0.050	0.043	0.050	0.368	0.274
	1.25	1.241	0.132	0.048	0.028	0.053	0.043	0.053	0.528	0.397
	1.30	1.299	0.138	0.060	0.030	0.058	0.044	0.058	0.672	0.543
	1.35	1.346	0.136	0.070	0.030	0.061	0.043	0.060	0.799	0.678
	1.40	1.396	0.144	0.079	0.031	0.068	0.043	0.067	0.880	0.776
	1.45	1.454	0.150	0.090	0.031	0.071	0.043	0.070	0.933	0.860
	1.50	1.497	0.158	0.099	0.033	0.075	0.042	0.073	0.966	0.899
1000	1.00	1.001	0.076	0.007	0.010	0.029	0.031	0.029	0.050	0.019
	1.05	1.047	0.080	0.013	0.014	0.030	0.031	0.030	0.097	0.073
	1.10	1.096	0.082	0.020	0.017	0.033	0.031	0.033	0.229	0.157
	1.15	1.148	0.085	0.030	0.019	0.036	0.030	0.036	0.452	0.329
	1.20	1.201	0.089	0.040	0.020	0.041	0.031	0.040	0.673	0.532
	1.25	1.247	0.091	0.049	0.021	0.045	0.030	0.044	0.839	0.695
	1.30	1.296	0.094	0.059	0.021	0.050	0.030	0.049	0.933	0.838
	1.35	1.352	0.101	0.071	0.022	0.055	0.030	0.054	0.975	0.920
	1.40	1.398	0.106	0.080	0.023	0.060	0.031	0.058	0.985	0.961
	1.45	1.444	0.104	0.090	0.022	0.066	0.031	0.064	0.996	0.987
	1.50	1.504	0.111	0.101	0.023	0.070	0.030	0.068	1.000	0.999
1500	1.00	1.001	0.060	0.005	0.008	0.024	0.026	0.024	0.039	0.023
	1.05	1.047	0.063	0.011	0.012	0.025	0.025	0.025	0.111	0.082
	1.10	1.102	0.070	0.021	0.015	0.028	0.025	0.028	0.347	0.267
	1.15	1.148	0.072	0.029	0.016	0.031	0.025	0.031	0.611	0.464
	1.20	1.199	0.073	0.040	0.017	0.036	0.025	0.036	0.841	0.690
	1.25	1.249	0.073	0.050	0.017	0.042	0.025	0.041	0.958	0.865
	1.30	1.302	0.080	0.061	0.018	0.047	0.025	0.047	0.990	0.950
	1.35	1.349	0.081	0.070	0.018	0.053	0.025	0.051	0.999	0.984
	1.40	1.403	0.083	0.081	0.018	0.058	0.025	0.056	1.000	0.994
	1.45	1.449	0.087	0.091	0.018	0.064	0.025	0.062	1.000	0.998
	1.50	1.499	0.086	0.100	0.018	0.068	0.025	0.066	1.000	1.000

doi:10.1371/journal.pone.0077399.t003

Note that under the IM model, HWE holds when $\rho = 0$, and HWE is not true when $\rho \neq 0$. Therefore, we propose the following LRT to test for haplotype-based HWE,

$$LRT_2 = 2ln\left[\frac{L(\hat{\mathbf{\Phi}})}{L(\hat{\mathbf{\Phi}}_0)}\right] = 2[lnL(\hat{\mathbf{\Phi}}) - lnL(\hat{\mathbf{\Phi}}_0)]$$

where $L(\hat{\mathbf{\Phi}}_0)$ and $L(\hat{\mathbf{\Phi}})$ are the values of the observed likelihood function under the null hypothesis of HWE and under the HWD alternative, respectively. Obviously, this LRT statistic asymptotically follows a Chi-square distribution with the degree of freedom being 1 when HWE holds.

Software Implementation

Based on the above EM, ECM and IEM algorithms, we have written a software HAP-HWE to conduct the proposed haplotype-based HWE tests, which is implemented in R (http://www.r-project.org) and is freely available at http://www.echobelt.org/web/UploadFiles/HAP-HWE.html. For each of the EM, ECM and IEM algorithms, let N denote the number of haplotypes that occur in all the possible haplotype combinations compatible with the observed genotypes $\mathbf{G}$ in the sample. As such, the initial values of all these N haplotype frequencies are taken as $1/N$ at $t = 0$. For the ECM and IEM algorithms, the initial values of K and ρ are taken as 1 and 0.01, respectively. The convergence criterion is that the absolute difference between the estimated values of the log-likelihood function at two consecutive iterations is smaller than

Table 4. Mean and standard deviation (SD) of K and ρ estimates, mean of sum of absolute differences (SAD) of haplotype frequency estimates for EM, ECM and IEM algorithms, simulated size and powers of two HWE tests for different values of ρ and n, under inbreeding model.

n	ρ	$\hat{K}$		$\hat{\rho}$		SAD			Size/Power	
		Mean	**SD**	**Mean**	**SD**	**EM**	**ECM**	**IEM**	**LRT$_1$**	**LRT$_2$**
500	0.00	1.001	0.111	0.010	0.015	0.041	0.044	0.041	0.060	0.020
	0.01	1.035	0.108	0.015	0.019	0.041	0.044	0.041	0.048	0.060
	0.02	1.077	0.120	0.024	0.023	0.041	0.044	0.041	0.114	0.137
	0.03	1.109	0.123	0.030	0.025	0.042	0.047	0.042	0.170	0.218
	0.04	1.157	0.124	0.041	0.027	0.041	0.048	0.041	0.268	0.345
	0.05	1.185	0.124	0.049	0.028	0.042	0.049	0.042	0.357	0.468
	0.06	1.241	0.131	0.061	0.028	0.041	0.051	0.041	0.528	0.629
	0.07	1.279	0.141	0.070	0.030	0.042	0.054	0.042	0.638	0.735
	0.08	1.316	0.144	0.079	0.031	0.043	0.057	0.043	0.718	0.811
	0.09	1.366	0.140	0.090	0.029	0.044	0.061	0.043	0.851	0.906
	0.10	1.409	0.156	0.099	0.032	0.043	0.064	0.043	0.889	0.940
1000	0.00	1.000	0.075	0.007	0.010	0.029	0.031	0.029	0.056	0.026
	0.01	1.032	0.076	0.012	0.014	0.029	0.030	0.029	0.053	0.077
	0.02	1.073	0.078	0.021	0.016	0.029	0.032	0.029	0.142	0.198
	0.03	1.115	0.084	0.030	0.018	0.029	0.034	0.029	0.294	0.394
	0.04	1.151	0.086	0.040	0.019	0.030	0.037	0.030	0.477	0.606
	0.05	1.188	0.088	0.049	0.020	0.030	0.039	0.030	0.643	0.762
	0.06	1.229	0.093	0.059	0.020	0.031	0.042	0.031	0.775	0.875
	0.07	1.276	0.097	0.070	0.021	0.029	0.045	0.029	0.895	0.957
	0.08	1.316	0.096	0.079	0.020	0.030	0.048	0.030	0.959	0.988
	0.09	1.357	0.102	0.089	0.021	0.030	0.052	0.030	0.973	0.995
	0.10	1.411	0.102	0.100	0.021	0.031	0.057	0.031	0.996	1.000
1500	0.00	1.001	0.064	0.006	0.009	0.024	0.025	0.024	0.053	0.028
	0.01	1.038	0.062	0.012	0.012	0.024	0.026	0.024	0.079	0.096
	0.02	1.073	0.066	0.020	0.014	0.024	0.027	0.024	0.210	0.272
	0.03	1.111	0.072	0.030	0.016	0.024	0.029	0.024	0.392	0.534
	0.04	1.150	0.070	0.040	0.015	0.024	0.032	0.024	0.610	0.763
	0.05	1.191	0.071	0.050	0.016	0.024	0.035	0.024	0.794	0.902
	0.06	1.231	0.075	0.059	0.017	0.025	0.038	0.025	0.916	0.970
	0.07	1.276	0.082	0.070	0.017	0.024	0.042	0.024	0.975	0.992
	0.08	1.318	0.081	0.080	0.017	0.025	0.047	0.025	0.995	0.997
	0.09	1.365	0.081	0.090	0.017	0.025	0.050	0.025	0.999	0.999
	0.10	1.408	0.086	0.100	0.018	0.025	0.054	0.025	1.000	1.000

doi:10.1371/journal.pone.0077399.t004

10^{-8}. The default maximum number of iterations is 1000. Then, the last estimates, $\hat{\mathbf{\Theta}}^{(t+1)}$, $\hat{\mathbf{\Psi}}^{(t+1)}$ and $\hat{\mathbf{\Phi}}^{(t+1)}$, are taken as the maximum likelihood estimates of $\hat{\mathbf{\Theta}}$, $\mathbf{\Psi}$ and $\mathbf{\Phi}$, respectively. Consequently, the values of LRT$_1$ and LRT$_2$ and the corresponding P values are obtained.

The input data file is a standard linkage pedigree file containing pedigree relationship, genotype and phenotype information, with each row being for an individual. The HAP-HWE software will only use the founders in the sample and automatically exclude the nonfounders from the analysis. Further, a haplotype block file is needed with each row representing a haplotype block, which can be easily exported from other existing software, such as Haploview [23]. Then, our HAP-HWE software will analyze the haplotype blocks one by one. The usage of the HAP-HWE software and other details refer to Text S2.

Our HAP-HWE software outputs: (i) the convergence processes of the log-likelihood function under the EM, ECM and IEM algorithms, (ii) the haplotypes with frequency estimates being larger than 10^{-5} and the associated frequency estimates under the three algorithms, (iii) the estimated value of K, the value of LRT$_1$ and the corresponding P value under the Niu's model, and (iv) the estimated value of ρ, the value of LRT$_2$ and the corresponding P value under the inbreeding model. The output results will be saved in a text file (named "results.txt") in the working directory. In addition, like other haplotype frequency estimation methods, our methods also face running time and

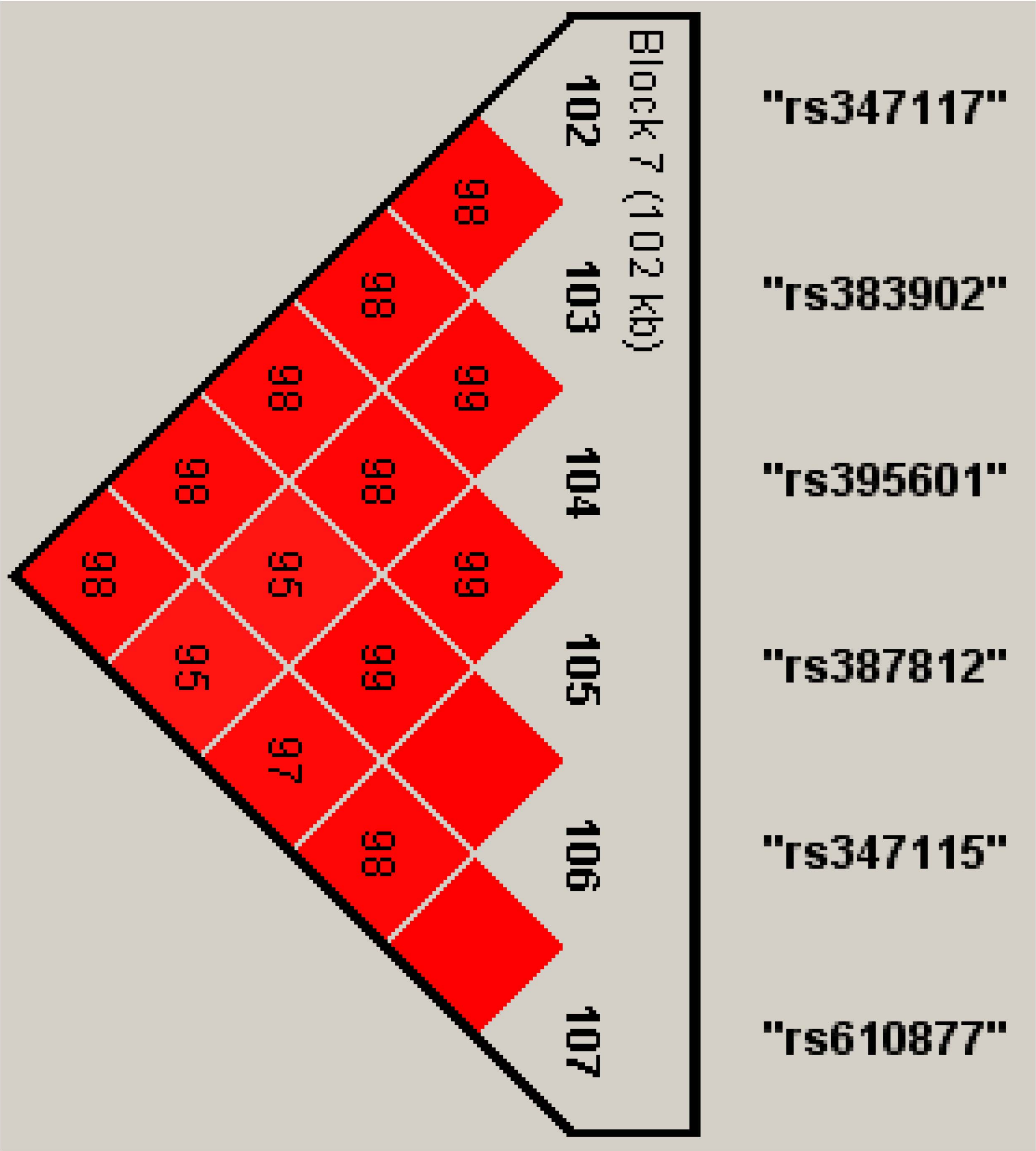

Figure 1. Haplotype LD display for the seventh haplotype block on chromosome 15. The red box denotes that the LOD value between any two loci is larger than or equal to 2.0. The numbers in the red boxes are the corresponding values of D' and the empty box denotes that $D'=1$.
doi:10.1371/journal.pone.0077399.g001

storage space problems because of the large number of possible haplotypes. In our software, to reduce storage space, each haplotype is represented by an integer, rather than a vector of alleles.

Results

Simulation Settings

To assess the validity and compare the performance of two LRT tests in testing for haplotype-based HWE, we consider three models with three tightly-linked SNPs that can lead to HWD:

Table 5. Mean and standard deviation (SD) of K and ρ estimates, mean of sum of absolute differences (SAD) of haplotype frequency estimates for EM, ECM and IEM algorithms, power comparison of two HWE tests under population stratification model, with the proportion τ of subpopulation I being taken as 0.6 and 0.8, and the sample size being fixed at 500, 1000 and 1500.

n	τ	$\hat{K}$		$\hat{\rho}$		SAD			Power	
		Mean	SD	Mean	SD	EM	ECM	IEM	LRT_1	LRT_2
500	0.6	1.224	0.144	0.044	0.022	0.073	0.094	0.072	0.396	0.572
	0.8	1.138	0.131	0.041	0.023	0.061	0.072	0.060	0.197	0.541
1000	0.6	1.223	0.103	0.045	0.016	0.061	0.086	0.061	0.674	0.874
	0.8	1.138	0.092	0.041	0.016	0.047	0.060	0.046	0.350	0.828
1500	0.6	1.226	0.083	0.045	0.013	0.057	0.084	0.056	0.850	0.962
	0.8	1.133	0.076	0.041	0.013	0.042	0.056	0.041	0.463	0.946

doi:10.1371/journal.pone.0077399.t005

Niu's model (NM), inbreeding model (IM) and population stratification (PS) model. For both the NM and IM models, the true marginal haplotype distribution is given in Table 1. For the NM model, the value of K is taken from 1.0 to 1.5 in increments of 0.05. Firstly, we calculate the probabilities of all the haplotype combinations from Equation (4). Then, one haplotype combination for each individual is randomly chosen. For the IM model, the inbreeding coefficient ρ is taken from 0 to 0.1 in increments of 0.01. Firstly, we calculate the probabilities of all the haplotype combinations from Equation (3), and then one haplotype combination is selected at random for each individual. Finally, we combine these two haplotypes to form the unphased genotype for the individual. To investigate how the population admixture affects the performance of two haplotype-based HWE tests, we consider the following PS model with two subpopulations I and II, where the corresponding haplotype distributions are given in Table 2, respectively. The proportion τ of the subpopulation I is taken to be 0.6 and 0.8.

Note that when $K=1$ and $\rho=0$, HWE holds for the NM and IM models, respectively. So, we simulate the type I error rates of the proposed HWE tests when $K=1$ or $\rho=0$, and make power comparison when $K>1$ and $\rho>0$. The PS model is also used to simulate the powers of both of the tests. For all the models, we generate samples of unrelated individuals at these three loci and the sample size is taken as 500, 1000 and 1500, respectively. The number of simulation replicates is fixed at 1000 and the significance level α is taken to be 5%.

As additional findings in this paper, we can also compare the efficiency of the EM, ECM and IEM algorithms in haplotype inference. The accuracy of haplotype frequency estimates is assessed by the sum of absolute differences (SAD) between the true and estimated frequencies, which was proposed by Fallin and Schork [20] and defined as

$$\mathrm{SAD}=\sum_{i=1}^{m}|\theta_i-\hat{\theta}_i|$$

where θ_i and $\hat{\theta}_i$ are the true and estimated haplotype frequencies of h_i, respectively. It ranges from 0 (when the estimation is perfect) to 1.

Simulation Results

Table 3 lists the estimate of K, mean SAD of haplotype frequency estimates, simulated size and powers of two HWE tests for different values of K and different sample sizes n under the Niu's model. It is shown in the table that the mean estimated value $\hat{K}$ over 1000 replicates is close to its true value. The type I error rate of LRT_1 is close to the nominal 5% level, while the size result of LRT_2 is less than 0.05, when $K=1$ (i.e. HWE holds). This means that in testing for haplotype-based HWE, LRT_1 controls the size well and LRT_2 is conservative under the NM model. The powers of both LRT_1 and LRT_2 are larger when K increases from 1.1 to 1.5 and the sample size n is fixed. However, LRT_1 is more powerful than LRT_2. In addition, when $K=1$ and n is unchanged, the EM, ECM and IEM algorithms perform similarly in the estimation of haplotype frequencies. However, with the increase of the K value, the SAD measure of the ECM algorithm does not have much change and is much smaller than the EM and IEM algorithms. The SADs of the EM and IEM algorithms are very close to each other and become larger when K is larger. On the other hand, with the sample size increasing, the SAD measures of all the three algorithms become less and two proposed LRT tests have more powers.

Table 4 shows the estimate of ρ, mean SAD of haplotype frequency estimates, simulated size and powers of two HWE tests for different values of inbreeding coefficient ρ and different sample sizes n under the inbreeding model. We can see from the table that the mean estimated value $\hat{\rho}$ over 1000 replicates is close to its true value. As shown in Table 3, LRT_1 performs better in controlling the size than LRT_2 under the IM model. However, LRT_2 is more powerful than LRT_1 under this situation. On the other hand, both the EM and IEM algorithms have the same performance and the corresponding SADs are stable across different values taken for ρ (0 to 0.1) in the estimation of haplotype frequencies. However, the ECM estimate gets larger with the increase of ρ and performs worse than the EM and IEM estimates. When the sample size is larger, the corresponding SADs appear to be smaller and two proposed LRT tests are more powerful.

Table 5 displays the mean SAD of haplotype frequency estimates and simulated powers of two HWE tests based on 1000 simulation replicates, under the PS model, with the proportion τ of subpopulation I being taken as 0.6 and 0.8, and the sample size being fixed at 500, 1000 and 1500. From the table, we find that LRT_2 is more powerful than LRT_1, irrespective of the τ value or the sample size n. In the estimation of haplotype frequencies, the EM and IEM algorithms perform similarly in SAD and have better SADs than the ECM estimate, which signifies that the EM and IEM algorithm are more robust to population stratification than the ECM algorithm.

Table 6. Results of application to North American Rheumatoid Arthritis Consortium data set.

Chr.	Haplotype block	N. of SNPs	SNP names	$\hat{K}$	$\hat{\rho}$	P-value	
						LRT$_1$	LRT$_2$
2	3	2	rs1686430, rs1734449	1.349	0.137	0.0107	0.0089
2	9	2	rs1866209, rs1438048	1.563	0.150	0.0051	0.0057
3	1	2	rs1516337, rs1516350	1.287	0.086	0.0204	0.0213
5	6	2	rs244903, rs244896	1.052	0.067	0.6341	0.0311
6	7	2	rs1565528, rs1491074	1.284	0.074	0.0231	0.0505
7	16	2	rs1182378, rs1182414	1.137	0.071	0.2526	0.0440
10	3	2	rs1979720, rs1494201	1.302	0.082	0.0189	0.0313
13	2	2	rs436227, rs390704	1.350	0.096	0.0085	0.0173
14	3	2	rs1381641, rs1020897	1.264	0.077	0.0326	0.0704
15	7	6	rs347117, rs383902, rs395601, rs387812, rs347115, rs610877	1.547	0.099	2.36×10^{-4}	0.0428
16	5	2	rs179209, rs179219	1.215	0.086	0.0708	0.0388
18	6	2	rs1787190, rs1981	1.330	0.139	0.0070	0.0076
21	3	2	rs1892687, rs2051179	1.372	0.146	0.0027	0.0048

Chr., chromosome; SNP, single nucleotide polymorphism; N. of SNPs, number of SNPs.
doi:10.1371/journal.pone.0077399.t006

Application to NARAC Data Set

We apply our HAP-HWE software to the Rheumatoid Arthritis (RA) data set from the North American Rheumatoid Arthritis Consortium (NARAC) [24], which was made available through the Genetic Analysis Workshop 15 [25]. In the data set, there are 757 pedigrees comprised of 8017 individuals (2481 founders and 5536 nonfounders), which were genotyped at 5407 SNP markers over the 22 autosomes. In each pedigree, there is at least one affected nonfounder with RA.

Note that information on haplotype blocks is needed prior to the HAP-HWE analysis. In this application, we use the existing software Haploview (version 4.2) [23] to define haplotype blocks, with all the arguments being taken as the default values. Then, 181 haplotype blocks are identified, 150 blocks including 2 SNPs, 19 blocks including 3 SNPs, 7 blocks including 4 SNPs, 1 block including 5 SNPs, 2 blocks including 6 SNPs, 1 block including 9 SNPs and 1 block including 13 SNPs.

On the other hand, HAP-HWE only uses the founders and excludes the nonfounders from the analysis. Further, there is a large proportion of missing genotypes for individuals in the data set. Therefore, the reduced data set used for the HAP-HWE analysis contains only a few founders in the data set. On the average, there are about 295 pedigrees (about 367 unrelated individuals) used for each haplotype block, ranging from 288 to 296 (ranging from 358 to 369).

Table 6 lists the results of the application to the NARAC data set. The significance level is fixed at $\alpha=5\%$. There are 13 haplotype blocks (out of 181) with at least one of the P values of the LRT$_1$ and LRT$_2$ being less than 5%. However, after multiple testing based on Bonferroni correction ($\alpha'=0.05/181=2.76\times10^{-4}$), only the seventh haplotype block including 6 SNPs (rs347117, rs383902, rs395601, rs387812, rs347115 and rs610877) on chromosome 15 is statistically significant with the P value of the LRT$_1$ being 2.36×10^{-4}. Figure 1 gives the Haploview LD display for this haplotype block. On the other hand, Min et al. [26] reported that chromosome 15p34 at rs347117 showed a possible linkage peak to RA by using the nonparametric linkage Z score ($Z=2.80$), which may support our finding.

Discussion

In this paper, we first wrote out two likelihood functions of the observed data based on the NM model and IM model. Then, we developed the ECM algorithm for the NM model to estimate the parameter K and haplotype frequencies and suggested the IEM algorithm for the IM model to estimate the inbreeding coefficient ρ and haplotype frequencies. Note that $K=1$ or $\rho=0$ means that HWE holds. So, we further proposed two LRT tests to test for haplotype-based HWE. We simulated the HWE, Niu's, inbreeding and population stratification models to assess the validity and compare the performance of these two LRT tests. The simulation results showed that both of the two tests are valid in testing for the haplotype-based HWE. If the Niu's model is true, then LRT$_1$ is more powerful. While, if the inbreeding model is true, then LRT$_2$ has better performance in power. Under the population stratification model, LRT$_2$ is still more powerful. Therefore, if the population model is unknown in practice, LRT$_2$ is generally recommended due to its good performance. Furthermore, we compared the performance of the EM, ECM and IEM algorithms in estimating the haplotype frequencies. If the true model is the Niu's model, then the ECM algorithm has more accurate estimates of haplotype frequencies than the EM and IEM estimates. However, for all the other simulation settings, the EM algorithm is not so much affected by the departure from HWE, and the EM and IEM algorithms almost have the same performance in controlling SAD, which is less than the ECM estimates. We also demonstrate the practical utility of the proposed methods by the application to the Rheumatoid Arthritis (RA) data set from the North American Rheumatoid Arthritis Consortium (NARAC). In addition, note that there are many abbreviations and notations used in this paper. So, in Supporting Information, we give two tables (Tables S1 and S2) to list them for the easy reference.

Supporting Information

Table S1 Summary of abbreviations.
(PDF)

Table S2 Summary of notations.
(PDF)

Text S1 Conditional-maximization steps of ECM algorithm.
(PDF)

Text S2 Help file of HAP-HWE.
(PDF)

Author Contributions

Conceived and designed the experiments: WGM JYZ. Performed the experiments: WGM HQH JYZ. Analyzed the data: WGM HQH YX PYC JYZ. Contributed reagents/materials/analysis tools: WGM HQH JYZ. Wrote the paper: WGM JYZ. Designed the software used in analysis: WGM JYZ. Revised the manuscript: YX PYC.

References

1. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. (2003) The international HapMap project. Nature 426: 789–796.
2. Gibbs RA, Belmont JW, Boudreau A, Leal SM, Hardenbol P, et al. (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
3. Zheng G, Yang Y, Zhu X, Elston RC (2012) Analysis of Genetic Association Studies. New York: Springer.
4. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544–548.
5. Huang BE, Amos CI, Lin DY (2007) Detecting haplotype effects in genomewide association studies. Genet Epidemiol 31: 803–812.
6. Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single- nucleotide polymorphisms. Am J Hum Genet 70: 157–169.
7. Yu Z, Schaid DJ (2007) Sequential haplotype scan methods for association analysis. Genet Epidemiol 31: 553–564.
8. Zhang K, Zhao H (2006) A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers from general pedigrees. Genet Epidemiol 30: 423–437.
9. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, et al. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67: 936–946.
10. Becker T, Knapp M (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27: 21–32.
11. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 1–38.
12. Excoffer L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921–927.
13. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425–434.
14. Stram DO, Pearce L, Bretsky P, Freedman M, Hirschhorn JN, et al. (2003) Modeling and EM estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55: 179–190.
15. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, et al. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53: 79–91.
16. Zhao LP, Li SS, Khalid N (2003) A method for the assessment of disease associations with singlenucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72: 1231–1250.
17. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73: 1316–1329.
18. Zeng D, Lin DY (2005) Estimating haplotype-disease associations with pooled genotype data. Genet Epidemiol 28: 70–82.
19. Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80: 267–278.
20. Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67: 947–959.
21. Kuk AYC, Zhang H, Yang Y (2009) Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium. Bioinformatics 25: 379–386.
22. Fan S (1989) A new extracting formula and a new distinguishing means on the one variable cubic equation (in Chinese). Natural Science Journal of Hainan Teachers College (in China) 2: 91–98.
23. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
24. Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, et al. (2003) Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. Arthritis Rheum 48: 906–916.
25. Amos CI, Chen WV, Remmers E, Siminovitch K, Seldin MF, et al. (2007) Data for Genetic Analysis Workshop (GAW) 15 Problem 2, genetic causes of rheumatoid arthritis and associated traits. BMC Proc (Suppl 1): S3.
26. Min JY, Min KB, Sung J, Cho SI (2010) Linkage and association studies of joint morbidity from rheumatoid arthritis. J Rheumatol 37: 291–295.