Published in final edited form as:

Stat Med. 2016 November 30; 35(27): 5040–5050. doi:10.1002/sim.7044.

A Composite Likelihood Approach in Testing for Hardy Weinberg Equilibrium using Family-Based Genetic Survey Data

Lingxiao Wang¹, Barry I Graubard², Yan Li^{1,2,*}

¹Joint Program in Survey Methodology, University of Maryland at College Park, MD 20742, USA

²Division of Cancer Epidemiology & Genetics, National Cancer Institute, Rockville, MD 20850, USA

Abstract

In population-based household surveys, for example, the National Health and Nutrition Examination Survey, households are often sampled by stratified multistage cluster sampling, and multiple individuals related by blood are often sampled within households. Therefore, genetic data collected from these population-based household surveys, called National Genetic Household Surveys, can be correlated due to two levels of correlation. One level of correlation is caused by the multistage geographical cluster sampling and the other is caused by biological inheritance among participants within the same sampled family. In this paper, we develop an efficient Hardy Weinberg Equilibrium (HWE) test utilizing pairwise composite likelihood methods that incorporate the sample weighting effect induced by the differential selection probabilities in complex sample designs, as well as the two-level clustering (correlation) effects described above. Monte Carlo simulation studies show that the proposed HWE test maintains the nominal levels, and is more powerful than existing methods (Li *et al.* 2011) under various (non)informative sample designs that depend on genotypes (explicitly or implicitly), family relationships, or both, especially when within-household sampling depends on the genotypes. The developed tests are further evaluated using simulated genetic data based on the Hispanic Health and Nutrition Survey.

Keywords

Complex Sampling; Condensed Coefficients of Identity; Intraclass Correlation; National Genetic Household Surveys; Taylor Linearization Variances

1. Introduction

Testing for Hardy Weinberg Equilibrium (HWE) for alleles of interest has been widely recommended as a crucial early step in genetic association studies. For example, two-stage testing procedures are frequently proposed [1], where testing HWE is the preliminary step before testing for association between the alleles and disease [2–4]. In addition, testing for HWE is often conducted before further testing for gene-environment (G-E) interactions in association with a disease as some G-E interaction testing approaches assume HWE [5].

*Correspondence: yli6@umd.edu.

National Genetic Household Surveys (NGHS) from various countries have begun to collect large-scale nationally representative samples of men and women, different racial/ethnic groups across age groups, with a wealth of exposure and behavioral information linked to DNA samples; examples are the Canadian Health Measures Survey [6], Health 2000 Survey from Finland [7], the U.S. National Health and Nutrition Examination Surveys [8] and the U.S. Health and Retirement Study [9–10]. Because these types of surveys select random samples of study subjects representing well-defined populations, they are less susceptible to ascertainment bias or other types of selection bias that can occur in traditional genetic studies, where a volunteer or convenience sample of families are selected. NGHS are rich in phenotypic, environmental, and behavioral data collected from medical examinations, personal interviews and linkage to medical records useful for conducting genetic association studies. The type of genetic data ranges, for example, from single-nucleotide polymorphisms (SNPs) of candidate genes in the NHANES III [8] with DNA collected in 1991–1994 to (planned) genome wide association data in the Health and Retirement Study [11].

Genetic data collected from NGHS are often correlated due to multistage geographical cluster sampling, as well as the possible biological inheritance among sample individuals within the household. For example in NHANES III, counties or cities are sampled at the first stage with nested cluster sampling at the later stages of sampling of segments and households. Within households, multiple individuals are often selected. In addition, sample weights that incorporate differential sampling rates, adjustments for non-response, and calibration to census totals, are used to obtain unbiased estimates of allele frequencies.

Methods of testing for HWE have been widely developed in simple random samples (SRS) [12–17]. Bourgain *et al.* [18] presented a testing method that accounted for genetic correlation among members within families while the families are selected with SRS. In the context of complex survey data where the families are selected with complex sampling, however, limited testing methods have been developed for testing for HWE.

Among the limited methods, tests developed by Li & Graubard [19] and Moonesinghe *et al.* [20] take account of the sample weighting and geographical clustering of complex survey sample designs. However, they did not specifically consider the clustering effect due to the biological inheritance within the family. She *et al.* [21] considered both levels of clustering for testing for HWE. However, their methods do not allow for differential selection probabilities within families, which, however, commonly occur in population-based household surveys. For example in NHANES III, adolescents aged 12–19 and persons with age 60+ years are sampled at higher rates within a sampled household. In addition, She *et al.* [21] only considered one type of familial relationship, i.e., two parents and one offspring. Li *et al.* [22] not only consider both levels of correlation, but also allow for differential selection probabilities of individuals within household. Their methods accommodated a variety of familial relationships. Nevertheless, the methods developed by Li *et al.* [22] produced inflated type I error when within-family sampling depended on disease status (or susceptible alleles implicitly).

To address issues of the complex survey sample design and within family biological correlation, we have developed a new modification of the HWE test, based on pairwise

composite likelihood and estimating equations methods, which utilize joint sample inclusion probabilities for each pair of individuals sampled within a family. This pairwise approach provides exact (not just asymptotic) unbiased estimation of the within family contribution to the estimating equation for any sample design where every pair of individuals within a sampled family have a calculable probability of being sampled. The proposed test of HWE maintains the nominal levels under various complex sample designs, including the design with within-family sampling dependent on disease status. The proposed method also addresses differential sampling rates at each stage of sampling i.e., the differential probabilities of sampling households and persons within the sampled households.

The statistical methodology is described in Section 2. We evaluate the performance of the new HWE test in terms of type I error rate and power via Monte Carlo simulation studies in Section 3. The proposed method is further evaluated using the Hispanic Health and Nutrition Survey (HHANES) [23] with simulated genotype data because HHANES provides detailed documentation about the within family sampling that allowed us to apply our proposed method.

2. Methods

Household interview surveys typically use geographical-base sampling where strata are initially formed consisting of large geographical units, such as counties, small cities or parts of large cities, which are called primary sampling units (PSU). Let there be H strata, with I_h PSUs randomly sampled from the h -th stratum at the first stage of sampling. Within sampled PSU- hi for $i = 1, 2, \dots, I_h$, data is collected on a random sample of J_{hi} families with K_{hij} individuals randomly sampled from the j th sampled family for $j = 1, 2, \dots, J_{hi}$. Consider a locus on a chromosome with a different alleles in the population. Let $M = a(a+1)/2$, the number of possible distinct genotypes. The data are represented by a vector of variables

$$y_{hijk} = (y_{hijk.1}, \dots, y_{hijk.g}, \dots, y_{hijk.M-1})$$

that are observed for each sampled individual, where $y_{hijk.g}$ equals to 1 if individual- $hijk$ has genotype g and 0 otherwise for $g = 1, \dots, M-1$. Let $E(y_{hijk}) = \mu_{hijk}$ with $\mu_{hijk} = (\mu_{hijk.1}, \dots, \mu_{hijk.g}, \dots, \mu_{hijk.M-1})^T$. If the genotype g is homozygote (e.g., I/I) then we assume the model $\mu_{hijk.g} = (1-r)p_I^2 + rp_I$, and if the genotype g is heterozygote (e.g., I/I' for allele I that is not allele I') then we assume $\mu_{hijk.g} = 2(1-r)p_I p_{I'}$ [24], where r is called fixation coefficient (i.e., the correlation between any pair of alleles within individuals).

Following Bourgain *et al.* [18] under the above model, we characterize the departure from HWE as the departure of r from zero. This model has a distinct parameters: the $(a-1)$ independent allele frequencies $p = (p_1, \dots, p_{a-1})^T$ and the fixation coefficient r with the following constraints on these parameters,

$$0 < p_l, \sum_{l=1}^a p_l = 1 \text{ and } \frac{-p_{\min}}{1-p_{\min}} \leq r \leq 1 \text{ with } p_{\min} = \min(p_1, \dots, p_a).$$

where the frequency of the last allele, $p_a = 1 - \sum_{l=1}^a p_l$. Under the null hypothesis of HWE, the fixation coefficient $r = 0$, and p is a nuisance parameter to be estimated.

The estimating equations for $\theta = (p, r)^T \theta$, assuming independence of the family members within each family, are given by

$$S(\theta) = \sum_{h=1}^H \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \sum_{k=1}^{K_{hij}} \left(\frac{\partial \mu_{hijk}}{\partial \theta} \right)^T w_{hijk} [\text{Var}(y_{hijk})]^{-1} (y_{hijk} - \mu_{hijk}) = 0, \quad (1)$$

where w_{hijk} is a $(M-1) \times (M-1)$ diagonal matrix with diagonal elements identically equal to the sample weight of individual- $hijk$ and $\text{Var}(y_{hijk})$ is the covariance matrix of y_{hijk} . The w_{hijk} are equal to the inverse of the probability that individual- $hijk$ is sampled from the target population; also typically the w_{hijk} include adjustments for nonresponses to the survey and poststratification to population census totals. The estimating equation (1) is derived under the assumption that the genotypes of sampled individuals are independent without considering that the genetic information conveyed by the family members can be correlated.

We relax the assumption that the genotypes of all family members within sampled family- hij are independent by defining the score function based on the pairwise likelihood [25], written as

$$S(\theta) = \sum_{h=1}^H \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} w_{hij} s_{hij} = 0,$$

where w_{hij} is the sample weight for sampled family- hij , which differs from the sample weight for individual- $hijk$ w_{hijk} that is defined earlier, and s_{hij} is the score for family- hij . The s_{hij} are estimated by the sum of scores over all pairs of family members, i.e.

$$\hat{s}_{hij} = \sum_{k=1}^{K_{hij}} \sum_{l=k+1}^{K_{hij}} \frac{1}{\pi_{kl|hij}} \left[\left(\frac{\partial \mu_{hij}}{\partial \theta} \right)^T [\text{Var}(y_{hij})]^{-1} (y_{hij} - \mu_{hij}) \right], \quad (2)$$

where $y_{hij} = (y_{hijk}, y_{hijl})^T$ and $\mu_{hij} = (\mu_{hijk}, \mu_{hijl})^T$ are vectors representing values for a pair of sampled family members (k, l) from the sampled family- hij , $\pi_{kl|hij}$ is the joint inclusion probability for selecting the pair (k, l) given that the family- hij is sampled, and $\text{Var}(y_{hij})$ is the covariance matrix of y_{hij} that accounts for dependence of family members and depends on the family relationship between the pair (k, l) ; see the definition for the pair of (parent, offspring) in equations (3)–(7) as an example.

To simplify the notation, we ignore the index of the j^{th} family in the i^{th} PSU in the h^{th} stratum, i.e. (hij) , and use y and μ to denote y_{hij} and μ_{hij} , respectively. Without loss of generality, consider the case of a biallelic locus (allele A and allele a , resulting in three genotypes AA , Aa and aa) and the family structure of two parents (P_1 and P_2) and two offspring (O_1 and O_2), and thus we have six pairwise relationships: P_1-P_2 , P_1-O_1 , P_1-O_2 , P_2-O_1 , P_2-O_2 , O_1-O_2 . For the P_1-O_1 , for example, we have

$$y = (y_{P_1, AA}, y_{P_1, Aa}, y_{O_1, AA}, y_{O_1, Aa})^T,$$

and its marginal mean is

$$\mu = ((1-r)p_A^2 + rp_A, 2(1-r)p_A(1-p_A), (1-r)p_A^2 + rp_A, 2(1-r)p_A(1-p_A))^T.$$

The first derivatives of μ with respect to p_A and r are, respectively,

$$\frac{\partial}{\partial p_A} \mu = (2(1-r)p_A + r, 2(1-r)(1-2p_A), 2(1-r)p_A + r, 2(1-r)(1-2p_A))^T$$

and

$$\frac{\partial}{\partial r} \mu = (1, -2, 1, -2)^T p_A(1-p_A).$$

For the construction of the covariance matrix $Var(y_{hij})$ in (2), we take account of the genetic correlation of the pair. The covariance matrix for the pair of P-O₁ in the family- *hij* can then be constructed as

$$Var(y) = \begin{bmatrix} \Sigma_P & \Sigma_{P, O_1} \\ SYS & \Sigma_{O_1} \end{bmatrix}, \quad (3)$$

where SYS denotes that $Var(y)$ is symmetric, and Σ 's are 2 by 2 covariance matrices between the indicators of genotypes in the same individual or between P and O₁. Following Bourgain *et al.* [18], the Σ 's are functions of the Condensed Coefficients of Identity (CCI) $= (1, 2, \dots, 9)^T$ for the nine condensed identity states and allele frequencies of p under the H_0 of $r=0$. Here CCI is determined by the family relationship between two individuals [26]. For example, if the sampled family consists of two parents (P_1 and P_2) and two full siblings (O_1 and O_2), then CCI between parent-offspring is

$$\Delta_{O_1}^{P_1} = \Delta_{O_2}^{P_1} = \Delta_{O_1}^{P_2} = \Delta_{O_2}^{P_2} = (\Delta_1 = \dots = \Delta_7 = 0, \Delta_8 = 1, \Delta_9 = 0)^T, \text{ between two full siblings is}$$

$$\Delta_{O_2}^{O_1} = (\Delta_1 = \dots = \Delta_6 = 0, \Delta_7 = \frac{1}{4}, \Delta_8 = \frac{1}{2}, \Delta_9 = \frac{1}{2})^T, \text{ and between two parents is}$$

$$\Delta_{P_2}^{P_1} = (\Delta_1 = \dots = \Delta_9 = 0)^T. \text{ Further calculation produces}$$

$$\Sigma_{P_1} = \Sigma_{P_2} = \Sigma_{O_1} = \Sigma_{O_2} = \begin{pmatrix} p_A^2 \cdot (1 - p_A^2) & -p_A^2 \cdot 2p_A p_a \\ SYS & 2p_A p_a \cdot (1 - 2p_A p_a) \end{pmatrix} \quad (4)$$

$$\Sigma_{P_1, O_1} = \Sigma_{P_1, O_2} = \Sigma_{P_2, O_1} = \Sigma_{P_2, O_2} = \begin{pmatrix} p_A^3 - p_A^4 & p_A^2 p_a - p_A^2 \cdot 2p_A p_a \\ SYS & p_A^2 p_a + p_a^2 p_A - (2p_A p_a)^2 \end{pmatrix} \quad (5)$$

$$\Sigma_{P_1, P_2} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad (6)$$

and

$$\Sigma_{O_1, O_2} = \begin{pmatrix} \frac{1}{4}p_A^2 + \frac{1}{2}p_A^3 + \frac{1}{4}p_A^4 - p_A^4 & \frac{1}{2}p_A^2p_a + \frac{1}{2}p_A^3p_a - p_A^2 \cdot 2p_Ap_a \\ SY S & \frac{1}{2}p_Ap_a + \frac{1}{2}(p_A^2p_a + p_Ap_a^2) + p_a^2p_A^2 - (2p_Ap_a)^2 \end{pmatrix} \quad (7)$$

In practice, relationships between two individuals in the family can vary, e.g., parent-offspring, half siblings, full siblings, etc.

To test the null hypothesis i.e. $H_0: r = 0$, we propose a quasi-score test statistic. Let $\tilde{\theta} = (\tilde{p}^w, r = 0)^T$ denotes the solution to $S_p(\tilde{\theta}) = 0$, where S_p is the first vector in the estimating equations (2) and $S(\theta) = (S_p^T, S_r^T)^T$ is partitioned in the same way as θ . Under suitable conditions [29], a quasi-score test statistic,

$$T_s^2 = \hat{S}_r^T(\tilde{\theta}) \tilde{V}_s^{-1} \hat{S}_r(\tilde{\theta}), \quad (8)$$

asymptotically follows a χ^2 distribution with 1 degree of freedom under the null hypothesis, where $\tilde{V}_s$ is a consistent estimator of $\text{Var}(\hat{S}_r)$ i.e. the covariance of $\hat{S}_r(\tilde{\theta})$. If the degrees of freedom for estimating $\text{Var}(\hat{S}_r)$ are not sufficiently large, $\tilde{V}_s^{-1}$ can be quite variable, which leads to the distribution of T_s^2 not being accurately approximated by a χ^2 distribution [29].

In this case the distribution of T_s^2 is approximated by an F-distribution with 1 degree of freedom in the numerator and $U-H$ degrees of freedom in the denominator under the null hypothesis, where $U = \sum_{h=1}^H I_h$ is the total number of sampled PSUs.

To estimate $\text{Var}(\hat{S}_r)$, a resampling method, such as jackknife or balanced repeated replication (BRR), is particularly attractive in the case of stratified multistage sampling because post-stratification and unit non-response adjustment can be automatically taken into account through the use of appropriate replicate weights [27]. For example, a jackknife estimator of $\text{Var}(\hat{S}_r)$ under stratified multistage sampling with I_h PSU's in the h^{th} stratum is given by

$$\hat{V}_J(\hat{S}_r) = \sum_{h=1}^H \frac{I_h - 1}{I_h} \sum_{i=1}^{I_h} \left(\hat{S}_r^{(hi)} - \hat{S}_r \right) \left(\hat{S}_r^{(hi)} - \hat{S}_r \right)^T.$$

Here $\hat{S}_r^{(hi)}$ is obtained in the same manner as $\hat{S}_r$ when the data from the cluster- (hi) is deleted, but using modified sample weight [28]. Alternatively, along the similar derivation discussed in Rao *et al.* [29], we can derive a Taylor linearization variance estimator (i.e., using a delta method)

$$\hat{V}_L(\hat{S}_r) = \sum_{h=1}^H \frac{I_h - 1}{I_h} \sum_{i=1}^{I_h} (z^{hi} - \bar{z}^h)(z^{hi} - \bar{z}^h)^T$$

where $z^{hi} = \sum_{j=1}^{J_{hi}} w_{hij} z_{hij}$ is the weighted PSU- hi total of $z_{hij} = s_{hij}^r - I_{21} I_{11}^{-1} s_{hij}^p$ (called Taylor deviates in the survey literature) with

$$s_{hij}^r = \sum_{k=1}^{K_{hij}} \sum_{l=k+1}^{K_{hij}} \frac{1}{\pi_{kl|hij}} \left\{ \left(\frac{\partial \mu_{hij}}{\partial r} \right)^T [Var(y_{hij})]^{-1} (y_{hij} - \mu_{hij}) \right\}, s_{hij}^p = \sum_{k=1}^{K_{hij}} \sum_{l=k+1}^{K_{hij}} \frac{1}{\pi_{kl|hij}} \left\{ \left(\frac{\partial \mu_{hij}}{\partial p} \right)^T [Var(y_{hij})]^{-1} (y_{hij} - \mu_{hij}) \right\},$$

$$I_{11} = -E \left[\frac{\partial}{\partial p} S_p(\theta) \right] = - \sum_{h=1}^H \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} w_{hij} E \left[\frac{\partial}{\partial p} s_{hij}^p \right],$$

$$I_{21} = -E \left[\frac{\partial}{\partial p} S_r(\theta) \right] = - \sum_{h=1}^H \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} w_{hij} E \left[\frac{\partial}{\partial p} s_{hij}^r \right],$$

$$E \left[\frac{\partial}{\partial p} s_{hij}^p \right] = \left(\frac{\partial \mu_{hij}}{\partial p} \right)^T [Var(y_{hij})]^{-1} \frac{\partial \mu_{hij}}{\partial p}, \text{ and } E \left[\frac{\partial}{\partial p} s_{hij}^r \right] = \left(\frac{\partial \mu_{hij}}{\partial r} \right)^T [Var(y_{hij})]^{-1} \frac{\partial \mu_{hij}}{\partial p},$$

evaluated at $\theta = \tilde{\theta}$ and $\bar{z}^h = \frac{1}{I_h} \sum_{i=1}^{I_h} z^{hi}$ is mean of the PSU totals in stratum h . An explicit assumption used for deriving $\hat{V}_L(\hat{S}_p)$ is that the sample of PSUs is approximately selected with replacement, which is an assumption used for estimating variances from multistage complex samples [28]. The estimator $\hat{V}_L(\hat{S}_p)$ is expressed as a function of the sample variances of the PSU-totals of Taylor deviates for each stratum. As a result, the correlations at the later sampling stages within the PSU, including the correlation among family members within the family, are captured in $\hat{V}_L(\hat{S}_p)$. Therefore, even if we do not account for the biological correlation within the family by using an independent working correlation matrix in Taylor deviate z^{hi} , the PSU-level variance estimators will still be robust and provide consistent variances that capture the correlations at the PSU-level of clustering and later stages of clustering. However, our quasi-score test statistic can achieve higher power by incorporating appropriate biological correlations within families through the working correlation matrix. Simulation studies in the next section support this anticipated increase in power.

3. Simulation Studies

We let the finite population consist of $M = 10,000$ PSUs with each PSU composed of 40 families. Considering a biallelic locus (allele A and allele a), the parental genotypes are generated independently according to a multinomial distribution with frequencies of

$P(aa) = (1 - r)p_a^2 + rp_a$, $P(Aa) = 2(1 - r)p_ap_A$, and $P(AA) = (1 - r)p_A^2 + rp_A$. Given the parental genotypes, the genotypes of the children are randomly generated according to Mendelian law. In order to introduce the clustering effect among families, we generate a clustered finite population. Specifically, we sort all of the 400,000 families in the population by the number of individuals with genotype *aa* within the family. The 10,000 PSUs are then formed by grouping every 40 families sequentially.

We randomly sample 400 PSUs from the total of 10,000 PSUs by employing two sampling designs: (1) simple random sampling (SRS) and (2) proportional to the population size sampling (PPS), where the sizes for Cluster 1–2000, Cluster 2001–4000, Cluster 4001–6000, Cluster 6001–8000, and Cluster 8001–10,000 are specified to be the values of 1, 2, 3, 4, and 5, respectively. As a result, in the PPS design, the sample weights depend on the genotypes and thus families with larger numbers of the *aa* genotype will be oversampled; while the sampled PSUs in the SRS design have equal selection probabilities. At the second stage, 10 out of 40 families within each sampled PSU are sampled with SRS.

For the sample selection of family members within sampled families at the third stage of sampling, three stratified SRS (SSRS) designs are employed and the stratum is, respectively, defined by 1) family relationship, denoted by SSRS(F); 2) both family relationship and genotypes, denoted by SSRS(GF). To ensure that each pair of family members has a positive inclusion probability, at least two family members need to be sampled if their corresponding stratum size ≥ 2 . Under SSRS(F), and SSRS(GF), family sizes are set to be 5, and 7, respectively, including the 2 parents. The computation of the joint inclusion probabilities (JIP) for each of the three SSRS design is shown in the Supplemental Material.

Under SSRS(F), there are two strata within each family: parent stratum and offspring stratum. We sample both parents and two out of three offspring by SRS. As a result, four family members are sampled from each of the 10 sampled families in the sampled cluster. Note SSRS(F) has sample design variables that is independent of the genetic variable of interest. For design SSRS(GF), we consider a case where the sample design depends on both the genetic variable of interest and a variable that is independent of it. Again, under this design, people with genotype *aa* are oversampled; see Supplemental Material Table A2 for further details.

The individual inclusion probability is the number of sampled family members divided by the number of total family members within strata. The JIP for each pair, however, depends on the stratum identity of the pair: if the pair comes from the same stratum, JIP is the number of sampled pairs divided by number of the total pairs; otherwise, JIP is the product of the two individual inclusion probabilities of the pair. Note that under SSRS(F), and SSRS(GF), each pair of family members has a positive probability of being sampled. Please see the computation of the joint inclusion weights, i.e. the reciprocal of the JIP, under SSRS(F), and SSRS(GF) in Supplemental Material Tables A1, and A3, respectively.

For comparison purposes, in addition to the proposed test statistic TS_p , we considered two existing test statistics TS_1 and TS_2 , described by Li *et al.* [22]. TS_1 incorporates both within-family biological correlation and between-family multistage cluster correlation. TS_2 only

incorporates the between-family correlation. When the sample design (in)directly depends on the genotypes, it was observed that TS_1 , the most powerful test under complex sample designs that are independent of genotypes, has inflated type I error.

To evaluate the performance of TS_p versus TS_1 and TS_2 , the rejection rate is calculated, i.e., the proportion of the 1,000 simulation runs for which the p-value is less than the nominal level α . When the fixation coefficient $r = 0$ ($r > 0$), the rejection rate is defined as size (power) of the test. In addition, two evaluation quantities are measured: (1) $\text{mean}(\hat{p}_a)$ i.e., mean of 1000 parameter estimates $\hat{p}_a$, and (2) $\text{VarRatio} = \text{Var}_1(\hat{S}_p) / \text{Var}_2(\hat{S}_p)$, where $\text{Var}_1(\hat{S}_p)$ is the mean of the 1000 proposed Taylor linearization estimates of the variances of $\hat{S}_p$ and $\text{Var}_2(\hat{S}_p)$ is estimated by the sample variance of the 1000 estimates $\hat{S}_p$. In the simulation studies, the value of p_a is varied at 0.1, 0.3 or 0.5. The results show similar patterns and therefore results at $p_a = 0.5$ only are presented in Tables 1–2 in the main text. Other results are available in Supplemental Material Tables A4–A6. Notice that when $p_a = 0.1$, the variance ratio increases when r gets bigger. This result is possibly due to the increasing correlation between the alleles as r increases, and therefore leading to reduced effective sample sizes. The proposed test, however, as r gets larger remains more powerful than TS_1 and TS_2 .

Table 1 presents the test size and the power of TS_p , TS_1 , and TS_2 under the SRS or PPS design for sampling clustered families, and the SSRS(F) design for sampling individuals within sampled families. The nominal level is specified at 5%. As expected, all three tests maintain nominal levels and TS_1 achieves highest power. For the SSRS(GF) design within families (Table 2), however, the allele frequency ($\text{mean}(\hat{p}_a)$) computed by TS_1 was biased and the corresponding type I errors (when $r=0$) were largely inflated. The other two tests, TS_p and TS_2 , which incorporate the first level of clustering, maintain the nominal level well. The proposed test TS_p is more powerful than TS_2 .

4. Example from the Hispanic Health and Nutrition Examination Survey

We use data from the Mexican-American part of the HHANES [23] that consisted of household interviews and physical examinations conducted between 1982 and 1984 for a random sample of non-institutionalized Mexican Americans aged 6 months to 74 years residing in sampled areas in the southwestern part of the US. The HHANES had a multistage stratified cluster sample design with $L = 8$ strata and 2 PSU's sampled per stratum i.e., the degrees of freedom for variance estimation is $U = 16 - 8 = 8$; see Gonzalez *et al.* [30] for further details about the sample design. At the last stage of sampling, individuals were sampled from participating households with rates based on age: 75% for 2–19 years, 50% for 20–44 years, and 100% for 45–74 years. This within-household sampling used a household roster obtained by the interviewer. The sample selection of each family member within sampled house is independent. Poisson sampling of individuals is approximated within households so that each pair of household members has a positive sample inclusion probability. Within a sampled household, all household members related by blood, marriage, or adoption were considered to be a family. To evaluate our methods using a real complex survey design, we restricted our analyses to families with married adults with or without children. Since our programs were set up for nuclear families that were no larger than five members sampled, we randomly sampled five individuals from each of the 67

families with greater than five members sampled. We adjusted their final weights by multiplying a factor of the number of sampled family members divided by five. In addition, since joint inclusion weight used in TS_p requires family size not smaller than 2, families with only one member selected are removed (see Discussion about a possible alternative approach that includes families with sample sizes of 1). As a result, there were 1,983 sampled families of which (761, 545, 416, or 261) of the sampled families had respectively (2, 3, 4, or 5) person(s) sampled. Thus, a total of 6,126 sampled individuals are involved in the data analysis. We adjusted the final weights so that the sum of the sampled individuals' final weights remain the same before and after removing families with only one member sampled. The final sample has eight strata with two PSUs nested within each stratum.

We applied the final sample weight for each sampled individual, w_{hijk} , which includes the inclusion probability at each stage of sampling and non-response and post-stratification adjustments. Since the HHANES did not genotype their sampled individuals, we generated genotype data using a two-step procedure. In step 1, the allele A frequency $p_A = 0.5$ is specified; in step 2, the parental genotypes are generated independently according to a multinomial distribution with frequencies of $p(aa) = (1 - r)p_a^2 + rp_a$, $p(Aa) = 2(1 - r)p_ap_A$, and $p(AA) = (1 - r)p_A^2$. Given parental genotypes, the genotype of a child was randomly generated according to Mendelian law.

The putative within-family inclusion probabilities $\pi_{k|hij}$ are 0.5, 0.75, or 1.0 in HHANES, if the k^{th} individual in family- (hij) has corresponding ages of <20 years, 20–44 years, or ≥45 years. Therefore, the within-family weight for sampling the k^{th} individual is 2 if the individual is < 20 years of old; 1.33 if the person is between 20–44 years of age; and 1 if the person is ≥45 years. We obtain the family-level sample weights by dividing each family member's final sample weight by the corresponding within-family weights (i.e. inverse of the putative within-family inclusion probabilities), and then taking the average [22]. By approximating within-family sampling as Poisson sampling, the conditional joint inclusion weight of the sampled pair (k^{th} and l^{th} individuals in family- (hij)) is the inverse of their conditional joint inclusion probability $\pi_{kl|hij}$, which is $\pi_{k|hij} \pi_{l|hij}$.

Table 3 presents results from the HHANES data analysis. Due to the small number of degrees of freedom for variance estimation, the test statistic is compared with F distribution with 1 degree of freedom in the numerator and 8 degrees of freedom in the denominator [29]. The results are consistent with the findings from simulation studies that the estimates of the allele frequency p_A are approximately unbiased and the proposed test TS_p reject the null hypothesis of $r=0$ when $r>0$. The larger value of r is, the smaller of the p-value is.

5. Conclusion and Discussion

Most conventional genetic studies can be classified into two categories: (1) population-based studies, in which unrelated individuals are sampled, (2) family-based studies, where households are sampled followed by blood related individuals within each family being sampled. In most genetic studies the sampling design for either category is specified and considered in the analysis, e.g., testing HWE. Apart from these two categories, this paper is motivated by population-based family data collected from NHGS, in which families are

sampled with a well specified complex sampling design, usually involving stratification and hierarchical geographic clustered sampling. As a result, there are two levels of clustering of the collected genetic data in NHGS. The first level of correlation is the hierarchical geographic clustered sampling of households; the second level of correlation is induced by biological inheritances from individuals sampled in the same household.

The genetic frequencies can be correlated within geographic clusters, which can lead to the inflation of variances of the estimates of these frequencies. Most national household surveys apply differential inclusion probabilities of sampling households and individuals within the sampled households across strata, usually defined by demographic or phenotypic characteristics, which can be correlated with the genetic markers of interest. Ignoring these inclusion probabilities in testing HWE can lead to biased estimates of genetic frequencies [22]. Therefore, these key sampling features need to be considered in our test methods in order to draw valid and efficient inferences for genetic data collected in NGHS.

Although motivated by the NGHS, the proposed test, which addresses all of the sample design complications, is not specific to NGHS. Instead, this test can have general application to common genetic studies such as family-based studies. Even though families are usually selected with SRS in these types of studies a researcher may want to do genetic testing on a subsample of family or pedigree members, where, for instance, the sampling rates may differ phenotypic characteristics. The proposed test could be used to efficiently take into account the differential sampling rates for testing HWE.

In this paper, we develop an efficient HWE test that considers two levels of clustering (correlation) effects with the first level induced by the hierarchical cluster sampling of families and the second level induced by the biological inheritance within the family under various complex designs for sampling families and members within families. Our simulation studies show that 1) when the within-family sampling only depends on genotypes, the developed test statistic, based on pairwise likelihood, yields unbiased estimates, maintains the nominal level and achieves the higher power; 2) when the within-family sampling is related to genotypes as well as family relationships, the advantage of the new test statistic is even more obvious in achieving higher test power.

The proposed test is developed for testing HWE in the general population from which the sample is randomly selected. Therefore, the proposed test would be applied to all the data regardless of the trait. In conventional case-control studies where the sampling fractions for the cases and controls are not used to weight the observations, the cases would be over represented in the combined sample of unweighted cases and controls, which would impact testing for HWE if the genes are related to the case status. Thus, using only the controls to test HWE, as is often done in convention case-control association studies, would closely reflect testing HWE in the population at risk. In population-based surveys where our approach weights the observations to reflect the underlying population, then using the entire sample of weighted cases and controls to test HWE is appropriate.

HWE is often assumed to achieve estimation efficiency in genetic association studies [31]. If HWE doesn't hold in the entire population, HWE are often tested within subpopulations,

commonly defined by race/ethnicity [19], followed by stratified association tests given HWE holds within subpopulations. Otherwise, a measure of the departure from HWE, for example, the fixation coefficient, can be used in the association test to adjust for a lack of HWE [32].

In conclusion, based on our small sample empirical results, the test statistics considering both levels of correlations should be used in population-based household surveys with multiple blood-related individuals sampled from the same household. Even if within-family sampling is related to both family relationships the genotypes, the new developed test statistic performs best. Please see R programs (an R package *svyhwe*) for simulations, real data analysis and a simple dataset in the Supplementary Material.

6. Limitation and Future Work

Even though the proposed method improves performance when within-family sampling is related to both family relationships and the genotypes, it is efficient only if there are at least two units, in other words, at least one pair of family members sampled in each household. If only one person was sampled within family, the pairwise estimating equation cannot be applied. In the data example using HHANES, families with only one member sampled were removed. To overcome this limitation, we suggest two possible solutions: (1) to improve the test statistic by combining the individual and pairwise test statistics; and (2) use a new (pseudo) likelihood function to develop a HWE test.

The first method creates a new quasi-score test statistic

$$T_{new}^2 = \begin{pmatrix} \hat{S}_r^I(\tilde{\theta}) \\ \hat{S}_r^P(\tilde{\theta}) \end{pmatrix}^T \cdot \begin{pmatrix} \tilde{V}_S^I & 0 \\ 0 & \tilde{V}_S^P \end{pmatrix}^{-1} \cdot \begin{pmatrix} \hat{S}_r^I(\tilde{\theta}) \\ \hat{S}_r^P(\tilde{\theta}) \end{pmatrix} \quad (9)$$

where $\hat{S}_r^I(\tilde{\theta})$ and $\hat{S}_r^P(\tilde{\theta})$ are individual and pairwise estimating equations. $\tilde{V}_S^I$ and $\tilde{V}_S^P$ are consistent estimators of $\text{Var}(\hat{S}_r^I(\tilde{\theta}))$ and $\text{Var}(\hat{S}_r^P(\tilde{\theta}))$ correspondingly. The advantage of this method is that all quantities in the test statistic are simple to calculate given that individual and pairwise estimating equations are well developed.

Another possible solution is to generate a new likelihood function combining likelihood functions applying individual weights and that using joint pair weights. This method requires completely new estimating equation, inferences and test statistic to be derived. A study of both approaches is planned in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Reference

1. Jahnes E, Muller DJ, Schulze TG, Windemuth C, Cichon S, Ohlraun S, Fangerau H, Held T, Maier W, Propping P, Nothen MM, Rietschel M. Association study between two variants in the DOPA

- Decarboxylase gene in bipolar and unipolar affective disorder. *Am J Med Genet.* 2002; 114:519–522. [PubMed: 12116187]
2. Salanti G, Amountza G, Ntzani EE, Ioannidis JP. Hardy-Weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. *Eur J Hum Genet.* 2005; 13:840–848. [PubMed: 15827565]
 3. Zou GY. Statistical methods for the analysis of genetic association studies. *Ann Hum Genet.* 2006; 70:262–276. [PubMed: 16626336]
 4. Zou GY, Donner A. The merit of testing Hardy-Weinberg Equilibrium in the analysis of unmatched case control data: A cautionary note. *Ann Hum Genet.* 2006; 70:923–933. [PubMed: 17044866]
 5. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika.* 2005; 92:399.
 6. Tremblay M, Wolfson M, Gorber SC. Canadian health measures survey: rationale, background and overview. *Health reports/Statistics Canada, Canadian Centre for Health Information.* 2007; 18:7–20.
 7. Heistaro S. Methodology report - Health 2000 Survey. 2007 Retrieved July 20, 2011
 8. NHANES. 2007. http://www.cdc.gov/nchs/nhanes/genetics/collection_dna.htm
 9. Health and Retirement Study, 1995. Technical description of the Health and Retirement Study Survey Sample Design.
 10. Health and Retirement Study. Sample Evolution: 1992–1998. 2008
 11. Health and Retirement Study, 2010. Genome Wide Association Study: A Resource for Genetic Research in the Behavioral and Health Sciences.
 12. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics.* 1992; 48:361–372. [PubMed: 1637966]
 13. Weir, BS. Genetic data analysis. Methods for discrete population genetic data. Sinauer Associates, Inc. Publishers; 1990.
 14. Ayres KL, Balding DJ. Measuring departures from Hardy-Weinberg: A Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity.* 1998; 80:769–777. [PubMed: 9705664]
 15. Shoemaker J, Painter I, Weir B. A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics.* 1998; 149:2079–2088. [PubMed: 9691059]
 16. Montoya-Delgado LE, Irony TZ, de B Pereira CA, Whittle MR. An unconditional exact test for the Hardy-Weinberg equilibrium law: Sample-space ordering using the Bayes factor. *Genetics.* 2001; 158:875–883. [PubMed: 11404348]
 17. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2005; 76:887–883. [PubMed: 15789306]
 18. Bourgain C, Abney M, Schneider D, Ober C, McPeck MS. Testing for Hardy-Weinberg Equilibrium in samples with related individuals. *Genetics.* 2004; 168:2349–2361. [PubMed: 15371359]
 19. Li Y, Graubard BI. Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data. *Biometrics.* 2009; 65:1096–1104. [PubMed: 19432778]
 20. Moonesinghe R, Yesupriya A, Chang MH, Dowling NF, Khoury MJ, Scott AJ. CDC/NCI NHANES III Genomics Working Group. A Hardy-Weinberg equilibrium test for analyzing population genetic surveys with complex sample designs. *Am J Epidemiol.* 2010; 171:932–941. [PubMed: 20237153]
 21. She D, Zhang H, Li Z. Testing Hardy-Weinberg equilibrium using family data from complex surveys. *Ann Hum Genet.* 2009; 73:449–455. [PubMed: 19489753]
 22. Li Y, Li Z, Graubard BI. Testing for Hardy Weinberg Equilibrium in National Household Surveys that Collect Family-Based Genetic Data. *Annals of human genetics.* 2011; 75:732–741. [PubMed: 21972963]
 23. NCHS: National Center for Health Statistics. Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982–84. *Vital and Health Statistics.* 1985; 1(19)
 24. Wright, S. The Theory of Gene Frequencies. Chicago, IL: The University of Chicago Press; 1969.
 25. Graubard BI, Korn EL. Conditional logistic regression with survey data. *Statistics in Biopharmaceutical Research.* 2011; 3:398–408.

26. Lange, K. Mathematical and Statistical Methods for Genetic Analysis. Springer Verlag; 2002. 82
27. Rust K, Rao JNK. Variance estimation for complex estimators in sample surveys. *Stat Med Res.* 1996; 5:381–397.
28. Korn, EL, Graubard, BI. Analysis of Health Surveys. New York: John Wiley & Sons, Inc.; 1999.
29. Rao JNK, Scott AJ, Skinner CJ. Quasi-score tests with survey data. *Stat Sin.* 1998; 8:1059–1070.
30. Gonzalez, JF, Ezzati, T, White, AA, Massey, T, Lago, J, Waksberg, J. Vital and Health Statistics. Vol. Series 1. Hyattsville, MD: National Center for Health Statistics; 1985. Sample design and estimation procedures. In Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982–84. DHHS Pub. No. (PHS) 85-1321
31. Chen J, Chatterjee N. Haplotype-based association analysis in cohort and nested case-control studies. *Biometrics.* 2006; 62:28–35. [PubMed: 16542226]
32. Schaid DJ, Jacobsen SJ. Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *Am J Epidemiol.* 1999; 149:706–711. [PubMed: 10206619]

Table 1

Results of 3 HWE tests for SRS and PPS sampling of clustered families with SSRS(F)-2P2O at 5% nominal level.

r^a	Test Statistics	SRS sampling			PPS sampling		
		$\text{mean}(\hat{\rho}_n)$	VarRatio^b	Rejection rate ^c	$\text{mean}(\hat{\rho}_n)$	VarRatio^b	Rejection rate ^c
0.00	TS ₁	0.500	1.013	0.06	0.498	1.053	0.05
	TS ₂	0.500	1.015	0.06	0.498	1.051	0.05
	TS _p	0.500	1.014	0.06	0.498	1.049	0.05
0.10	TS ₁	0.500	0.991	0.73	0.499	1.075	0.76
	TS ₂	0.500	0.987	0.59	0.499	1.071	0.59
	TS _p	0.500	0.989	0.65	0.499	1.073	0.67
0.15	TS ₁	0.500	1.018	0.98	0.499	1.062	0.99
	TS ₂	0.500	1.012	0.88	0.499	1.063	0.92
	TS _p	0.500	1.014	0.94	0.498	1.063	0.97
0.20	TS ₁	0.500	1.067	1.00	0.498	1.031	1.00
	TS ₂	0.500	1.053	0.99	0.498	0.999	0.99
	TS _p	0.500	1.054	1.00	0.498	1.016	1.00

^a r = fixation coefficient

^b $\text{VarRatio} = \text{var}1(\hat{S}_P) / \text{var}2(\hat{S}_P)$, where $\text{var}1(\hat{S}_P)$ is the mean of the 1,000 estimated variances of $\hat{S}_P$ and $\text{var}2(\hat{S}_P)$ is estimated by the sample variance of the 1,000 estimates $\hat{S}_P$

^c Rejection rate under the hypothesis of fixation coefficient $r = 0$ (> 0) is the test size (power).

Table 2

Results of 3 HWE tests for SRS and PPS sampling of clustered families with SSRS(GF) at 5% nominal level.

r^a	Test Statistics	SRS sampling			PPS sampling		
		$\text{mean}(\hat{\rho}_n)$	VarRatio^b	Rejection rate^c	$\text{mean}(\hat{\rho}_n)$	VarRatio^b	Rejection rate^c
0.00	TS ₁	0.410	1.248	0.46	0.409	1.119	0.54
	TS ₂	0.500	1.046	0.04	0.498	1.033	0.04
	TS _P	0.500	1.047	0.04	0.498	1.033	0.04
0.10	TS ₁	0.415	1.177	0.05	0.412	1.014	0.06
	TS ₂	0.500	1.015	0.27	0.498	0.987	0.27
	TS _P	0.500	1.018	0.34	0.498	0.989	0.34
0.15	TS ₁	0.416	1.255	0.06	0.414	1.011	0.10
	TS ₂	0.500	1.086	0.54	0.498	1.025	0.55
	TS _P	0.500	1.089	0.66	0.498	1.026	0.66
0.20	TS ₁	0.418	1.194	0.26	0.416	0.996	0.32
	TS ₂	0.500	1.046	0.76	0.498	1.039	0.77
	TS _P	0.500	1.051	0.86	0.498	1.040	0.87

^a r = fixation coefficient

^b $\text{VarRatio} = \text{var}1(\hat{S}_P) / \text{var}2(\hat{S}_P)$, where $\text{var}1(\hat{S}_P)$ is the mean of the 1,000 estimated variances of $\hat{S}_P$ and $\text{var}2(\hat{S}_P)$ is estimated by the sample variance of the 1,000 estimates $\hat{S}_P$

^c Rejection rate under the hypothesis of fixation coefficient $r = 0$ (> 0) is the test size (power).

Table 3

Results from the Hispanic Health and Nutrition Examination Survey data analysis where the genotype data are generated with allele frequency $p_a = 0.5$ and varying fixation coefficient r

Test Statistic	$\hat{p}_a$			P-values		
	$r = 0$	$r = 0.05$	$r = 0.1$	$r = 0$	$r = 0.05$	$r = 0.1$
TS ₁	0.50	0.50	0.50	0.81	0.02	<0.01
TS ₂	0.50	0.49	0.50	0.33	0.02	<0.01
TS _p	0.50	0.50	0.50	0.35	0.01	<0.01