Published in final edited form as:

Stat Med. 2008 December 10; 27(28): 5801–5815. doi:10.1002/sim.3407.

Testing linkage disequilibrium from pooled DNA: a contingency table perspective

Jinfeng Xu¹, Yaning Yang^{*2,†}, Zhiliang Ying³, and Jurg Ott^{4,5}

¹Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore

²Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China

³Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

⁴Lab of Statistical Genetics, Rockefeller University, New York, NY 10021, U.S.A.

⁵Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.

SUMMARY

Pooling DNA samples of multiple individuals has been advocated as a method to reduce genotyping costs. Under such a scheme, only the allele counts at each locus, not the haplotype information, are observed. We develop a systematic way for handling such data by formulating the problem in terms of contingency tables, where pooled allele counts are expressed as the margins and the haplotype counts correspond to the unobserved cell counts. We show that the cell frequencies can be uniquely determined from the marginal frequencies under the usual Hardy-Weinberg equilibrium (HWE) assumption and that the maximum likelihood estimates of haplotype frequencies are consistent and asymptotically normal as the number of pools increases. The limiting covariance matrix is shown to be closely related to the extended hypergeometric distribution. Our results are used to derive Wald-type tests for linkage disequilibrium coefficient using pooled data. It is discovered that pooling is not efficient in testing weak linkage disequilibrium despite its efficiency in estimating haplotype frequencies. We also show by simulations that the proposed linkage disequilibrium tests are robust to slight deviation from HWE and to minor genotype error. Applications to two real angiotensinogen gene data sets are also provided.

Keywords

Contingency table; DNA pooling; haplotype; hypergeometric distribution; linkage disequilibrium

1. INTRODUCTION

For large scale genetic studies, pooling multiple individual DNA samples is used to reduce genotyping cost [1]. A common difficulty in pooling design is that the haplotypes are not directly observable with most of the current genotyping platforms. A single nucleotide polymorphism (SNP) haplotype is a sequence of SNP alleles on one chromosome and is thought to be informative in differentiating populations. Haplotype-based association

analysis is an important tool for detecting gene-disease associations [1]. Estimation of haplotype frequencies and haplotype-based association analysis from pooled DNA data was studied by several authors [2–4]. However, linkage disequilibrium (LD) test for pooled DNA data has not been investigated. In this paper, from a contingency table perspective, we derive two Wald-type tests of LD. Considering that linkage disequilibrium is of primary interest in gene mapping, performances of different pooling schemes (e.g. varying pool sizes) can be compared by power of LD test.

Haplotype analysis for pooled DNA poses a challenge because both phase information and individual information are lost. With family data, phase information may be recovered by examining genotypes from both parents and offsprings. For population data, the Hardy-Weinberg equilibrium (HWE) [5,6] is required to be true. Under the HWE, the genotype observations of a pool of K unrelated individual DNA samples can be viewed as a pool of $2K$ independent haplotypes. We shall call a pool of K individuals as a K -pool. Note that even for $K = 1$, the genotype is in fact a pool of two haplotypes. For this reason, we will henceforth regard an individual DNA sample as a 1-pool and call a haplotype a $\frac{1}{2}$ -pool. In other words, an 1-pool corresponds to an individual and an $\frac{1}{2}$ -pool corresponds to a single haplotype.

Estimating haplotype frequencies from pooled genotype can be viewed as estimating cell frequencies of a contingency table with only certain marginal counts being observed. We show that under the HWE assumption, the cell frequencies are uniquely determined by the marginal frequencies. We obtain large sample properties based upon such a formulation. In fact, the distribution of cell counts with fixed margins follows the family of well-known extended hypergeometric distributions [7]. In this connection, the approximations and computing algorithms for the mean and variance of the univariate noncentral hypergeometric distributions as developed in [7,8,9] can be used. In addition, by using the contingency table formulation, we are able to characterize the relationship of the variance of the estimated haplotype frequencies with pool sizes, which were not derived in [4]. Based on these results, we derive two Wald-type LD tests for pooled data. These tests can also be applied to individual DNA data. When two loci are in linkage equilibrium, the variance of estimated LD coefficient has an explicit analytic form and the efficiency of the LD test is shown to be independent of pool sizes.

The paper is organized as follows. Section 2 discusses the contingency table formulation of the haplotype estimation problem. By utilizing its close connection to the extended hypergeometric distribution in the case of two-loci, we obtain in Section 3 an explicit variance formula which can be evaluated accurately via simple Monte-Carlo simulation. Section 4 covers estimation and Wald-type tests for linkage disequilibrium coefficient. Theoretical calculation shows that pooling is not efficient in testing linkage disequilibrium when linkage disequilibrium is weak. Section 5 conducts simulation studies to investigate efficiency of pooling in testing linkage disequilibrium. This section also investigates robustness of the proposed tests when HWE is slightly violated or when there is genotype error. Section 6 applies our methods to single-nucleotide polymorphisms in the angiotensinogen gene. Section 7 contains discussion and some concluding remarks.

2. CONTINGENCY TABLE CONNECTION

Suppose that DNA samples from K unrelated individuals are pooled. The case of $K = 1$ corresponds to a single individual DNA sample whereas $K = \frac{1}{2}$ corresponds to a single haplotype. In general, each pool contains $2K$ haplotypes which are not observed directly. Without loss of generality, the two alleles at each locus are denoted by 0 and 1. Thus, a haplotype of m loci is an m -vector of 0's and 1's, i.e. any of the following 2^m haplotypes:

$$H_1=(0, 0, \dots, 0)^T, H_2=(1, 0, 0, \dots, 0)^T, \dots, H_{2^m}=(1, 1, 1, \dots, 1)^T.$$

Let $h_j = \text{pr}\{H_j\}$ be the relative frequency (probability) of the j -th haplotype, $j = 1, \dots, 2^m$ and write $\mathbf{h} = (h_1, h_2, \dots, h_{2^m})^T$. The observed genotype at the m loci is denoted by $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$, where x_k is the number of allele 1 at the k th locus while $y_k = 2K - x_k$ corresponds to the number of allele 0. Obviously, both x_k and y_k are integers bounded by 0 and $2K$.

Using each of the m loci as a classifier, we can treat each DNA pool as a multi-way two-level contingency table, cross-classifying $2K$ independent chromosomes by their haplotypes. However, because of pooling, only the genotypes, i.e., the margins of the table are observed. The main objective is to estimate cell or haplotype frequencies from the observed marginal counts.

Let us first consider the case of $m = 2$. We will use ξ_1 to denote the number of haplotype (0, 0), and, likewise, use ξ_2, ξ_3 and ξ_4 for haplotypes (0, 1), (1, 0) and (1, 1), respectively. Define $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)^T$. Note that ξ is not directly observable. Since each allele takes value 1 or 0, the two loci cross-classify the data as a 2×2 table given below.

		locus 2		
		0	1	
locus 1	0	ξ_1	ξ_2	y_1
	1	ξ_3	ξ_4	x_1
		y_2	x_2	$2K$

The general case of m -loci corresponds to a 2^m -table. It is clear that cell frequencies in a contingency table are generally not uniquely determined by the marginal frequencies. The identifiability in the present setting is due to the HWE assumption, as to be shown below. Let ξ_k be the unobserved cell counts for haplotype H_k and $\xi = (\xi_1, \dots, \xi_{2^m})^T$. Note that for a given haplotype configuration J of a pool, ξ_k is the number of haplotype k in the collection of J . Due to the HWE assumption, the probability mass function, $p_{\mathbf{h}}(\xi)$, of ξ is clearly a multinomial distribution with probabilities $\mathbf{h}$ and total number $2K$. Then, the probability mass function of $\mathbf{x}$ is given by

$$p_{\mathbf{h}}(\mathbf{x}) = \sum_{\xi \in H_{\mathbf{x}}} p_{\mathbf{h}}(\xi) = \sum_{\xi \in H_{\mathbf{x}}} \binom{2K}{\xi_1, \xi_2, \dots, \xi_{2^m}} h_1^{\xi_1} h_2^{\xi_2} \dots h_{2^m}^{\xi_{2^m}},$$

where $H_{\mathbf{x}}$ is the collection of all cell counts that are consistent with the pool genotype $\mathbf{x}$ (the y_k will be ignored since $y_k = 2K - x_k$, $k = 1, \dots, m$).

Assume without loss of generality that $h_j > 0$, $j = 1, \dots, 2^m$ and denote all possible values of $\mathbf{x}$ by $\chi = \{0, \dots, 2K\}^m$. We now show that the marginal probability mass function $\{p(\mathbf{x}), \mathbf{x} \in \chi\}$ uniquely determines $\{h_j = \text{pr}(H_j), j = 1, \dots, 2^m\}$, the probability mass function for the haplotypes. For any H_j , let $\mathbf{x}(j) = 2KH_j$, which has 0 or $2K$ as its components. It is clear that, to attain this configuration, all haplotypes must be H_j . Therefore,

$p_{\mathbf{h}}(\mathbf{x}(j)) = h_j^{2K}$ or $h_j = p_{\mathbf{h}}^{\frac{1}{2K}}(\mathbf{x}(j))$. Hence, the cell probabilities are identifiable from the margins of a sequence of 2×2 tables in this case.

Now, for n independent and identically distributed such contingency tables with

observations $\mathbf{x}_i$, the log-likelihood is $\sum_{i=1}^n \log p_{\mathbf{h}}(\mathbf{x}_i)$. The smoothness of the likelihood function and the identifiability of $\mathbf{h}$ implies the consistency and asymptotic normality of the maximum likelihood estimate [10]. Denote $E\{-\frac{\partial^2 \log p_{\mathbf{h}}(\mathbf{x})}{\partial \mathbf{h} \partial \mathbf{h}^T}\}$ by $I_K(h)$ and take the constraint on $\mathbf{h}$, $\sum_{l=1}^{2^m} h_l = 1$, into consideration [11,4],

$$n^{\frac{1}{2}}(\hat{\mathbf{h}} - \mathbf{h}) \rightarrow N(0, \sum_K)$$

in distribution, where the variance-covariance matrix $\sum_K$ is given by

$$\sum_K = W\{W^T I_K(\mathbf{h}) W\}^{-1} W^T, \quad (1)$$

where $W = (I_{2^m-1}, -\mathbf{1})^T$ is a $2^m \times (2^m - 1)$ matrix, I_{2^m-1} is the $(2^m - 1) \times (2^m - 1)$ identity matrix, $\mathbf{1}$ is the vector of 1's of length $2^m - 1$, and $\mathbf{a}^{\otimes 2} = \mathbf{a} \mathbf{a}^T$ for a vector $\mathbf{a}$. Using latent variable ξ , we show in the Appendix A that $I_K(h)$ has the following simple expression

$$I_K(\mathbf{h}) = 2K\Gamma - \Gamma E\{\text{var}(\xi|\mathbf{x})\}\Gamma. \quad (2)$$

where Γ is the diagonal matrix with diagonal elements $\Gamma(j, j) = 1/h_j$, $1 \leq j \leq 2^m$. This expression links the asymptotic variance-covariance matrix with the latent multinomial variables ξ and allows further investigations through tools of contingency table.

Conditioning on the margins of a contingency table, the cell counts have an m -way extended hypergeometric distribution. Hence introduction of the unobserved ξ connects the problem to the well-known hypergeometric distribution, which is widely used in contingency table analysis. In the next section, we show that this connection also facilitates detailed investigations of the limiting covariance matrix and efficiency of LD tests with pooled data.

3. PROPERTIES OF THE LIMITING COVARIANCE MATRIX

In view of (1) and (2), in order to study properties of the limiting covariance matrix $\sum_K$, we need to evaluate $E\{\text{var}(\xi|\mathbf{x})\}$. In general, it is not clear how to explore the conditional distribution of ξ given $\mathbf{x}$. Throughout the rest of this paper, we will confine ourselves to the special case of $m = 2$, which is important as very often linkage disequilibrium are studied pairwise [12,13]. In this case, conditional on $\mathbf{x}$, ξ has an extended hypergeometric distribution [7]. Except for the special case of central hypergeometric distribution, there does not exist simple variance formula. However, fast and accurate numerical procedures do exist for computing variance [8].

Recall that for $m = 2$, $\mathbf{x} = (x_1, x_2)^T$ and $y_k = 2K - x_k$ ($k = 1, 2$), $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)^T$ and $\mathbf{h} = (h_1, h_2, h_3, h_4)^T$. Note that $\xi \sim M(2K, \mathbf{h})$. Conditional on $\mathbf{x}$ or, equivalently, on x_1, y_1, x_2, y_2 , the ξ_1 has an extended hypergeometric probability mass function

$$\text{pr}\{\xi_1|x_1, x_2\} = \frac{\binom{y_1}{\xi_1} \binom{x_1}{y_2 - \xi_1} \psi^{\xi_1}}{\sum_j \binom{y_1}{j} \binom{x_1}{y_2 - j} \psi^j},$$

where $\psi = h_1 h_4 / (h_2 h_3)$ is the odds ratio. Since the four components of ξ are perfectly correlated given $\mathbf{x}$, it follows that

$$\begin{aligned} \text{cov}(\xi|\mathbf{x}) &= \text{var}(\xi_1|\mathbf{x}) \mathbf{u}_1^{\otimes 2}, \\ \text{where } \mathbf{u}_1 &= (1, -1, -1, 1)^T. \text{ Letting } v_K = E\{\text{var}(\xi_1|\mathbf{x})\}, \text{ we have} \\ E\{\text{cov}(\xi|\mathbf{x})\} &= E\{\text{var}(\xi_1|\mathbf{x})\} \mathbf{u}_1^{\otimes 2} = v_K \mathbf{u}_1^{\otimes 2}. \end{aligned} \quad (3)$$

Making use of (3), we prove in the Appendix B the following proposition.

PROPOSITION

For $m = 2$, the limiting covariance matrix has the following expression

$$\Sigma_K = \frac{1}{2K} \left\{ \text{diag}(\mathbf{h}) - \mathbf{h}^{\otimes 2} \right\} + \frac{v_K}{2K \left\{ 2K - v_K (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}) \right\}} \mathbf{u}_1^{\otimes 2}. \quad (4)$$

The first term on the right-hand side of (4) corresponds to the maximum likelihood estimate when all $2K$ haplotypes are observed. The second term therefore represents the additional variation due to pooling. It involves v_K which is a function of odds ratio ψ . Σ_K measures precision of estimated haplotype frequencies and decrease of Σ_K with pool size K shows efficiency of pooling in haplotype estimation. By simulation studies, Yang et al. [4] found that pooling is efficient when two loci are strongly associated. It is interesting to ask whether pooling is still efficient when two loci are independent, i.e. in linkage equilibrium. Denote v_K

and $\sum_K$ by v_K^1 and $\sum_K$ respectively in the special case of $\psi = 1$, we can show that $v_K^1 = (2K - 1) / (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})$. Making use of it, we prove in the Appendix C the following corollary which indicates efficiency of pooling in haplotype estimation when two loci are in linkage equilibrium.

COROLLARY 1

When $\psi = 1$ and $m = 2$,

$$\sum_K^1 = (2K)^{-1} \left\{ \sum_{\frac{1}{2}} - (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})^{-1} \mathbf{u}_1^{\otimes 2} \right\} + (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})^{-1} \mathbf{u}_1^{\otimes 2}.$$

In particular, $\sum_K^1$ decreases monotonically to $(h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})^{-1} \mathbf{u}_1^{\otimes 2}$ as K increases to ∞ .

The expression for $\sum_K^1$ consists of two terms only, the first of which involves the pool size and contributes less as pool size increases. Because the rank of $\mathbf{u}_1^{\otimes 2}$ is 1, the limit of $\sum_K^1$

must be singular with rank = 1. Therefore, components of $\hat{h}$ must be perfectly correlated as $K \rightarrow \infty$. This can be explained naturally in terms of the constraints on h in the limiting case. For $K = \infty$, all the marginal probabilities can be estimated without error and thus regarded as fixed. With the margins fixed, there is only one degree of freedom for vector $\mathbf{h}$. In other words, the value of any component in $\mathbf{h}$ determines the rest.

In the special case of linkage equilibrium, Corollary 1 obtains the closed-form variance formula. The variance for cases with general LD can be calculated by Monte Carlo method. Specifically, one can generate a large number of random variables from *Multinomial*($2K$, (h_1, h_2, h_3, h_4)), calculate conditional variance $\text{var}(\xi_1|x)$ using Liao and Rosen's algorithm [8] for each Monte Carlo replicate, then average over these conditional variances to get an approximation of v_K .

4. WALD-TYPE TESTS FOR LINKAGE DISEQUILIBRIUM

Pfeiffer et al. [14] studied the problem of estimating LD coefficients from qualitative assays for carrier status of risk allele in pooled samples. Here we consider LD test for qualitative assays of pooled genotypes. Define $D = h_1 - (h_1 + h_2)(h_1 + h_3) = h_1h_4 - h_2h_3$, which is the so-called linkage disequilibrium coefficient. This is an important quantity in population genetics and plays a critical role in determining the power of association studies. Since it is a function of haplotype frequencies $\mathbf{h}$, its maximum likelihood estimate must be $\hat{D} = \hat{h}_1\hat{h}_4 - \hat{h}_2\hat{h}_3$. Although estimation of linkage disequilibrium coefficient is straightforward, test for linkage disequilibrium coefficient with pooled data has not been formally addressed.

In this section, we derive Wald-type tests for linkage disequilibrium coefficient which have simple forms and also gain us many insights into efficiency of pooling. Although we derive the Wald-type tests for general pool size K , the special case of $K = 1$ may be of particular interest since individual genotype data are most commonly used in practical applications.

By the δ -method [15] and the asymptotic normality of $\hat{\mathbf{h}}$, we can show that $\sqrt{n}(\hat{D} - D)$ converges in distribution to normal with mean 0 and variance $B_K(h)$ equal to

$$\frac{h_1h_4(h_1+h_4)+h_2h_3(h_2+h_3)-4(h_1h_4-h_2h_3)^2}{2K} + \frac{(2K)^{-1}v_K}{2K - v_K(h_1^{-1}+h_2^{-1}+h_3^{-1}+h_4^{-1})}.$$

In the case of linkage equilibrium, i.e. $D = 0$ or $\psi = 1$, the above variance formula can be further simplified. Specifically, denote $B_K(h)$ by $B_K^0(h)$ when $D = 0$, using

$v_K^1 = (2K - 1)/(h_1^{-1}+h_2^{-1}+h_3^{-1}+h_4^{-1})$, we can show that $B_K^0(h) = (h_1^{-1}+h_2^{-1}+h_3^{-1}+h_4^{-1})^{-1}$. This is rather surprising because the expression does not depend on the pool size K . Consequently, it indicates that the method of pooling does not have much effect in terms of efficiency gain when the linkage disequilibrium is weak.

The preceding results can be readily used to obtain a Wald-type test for $H_0 : D = D_0$ versus $H_a : D \neq D_0$. Specifically, we define D test as

$$\frac{\sqrt{n}(\hat{D} - D_0)}{\sqrt{B_K(\hat{h})}}$$

and use the standard normal to get the critical region. Note that testing for no LD (i.e., $D_0 = 0$) is of general interest in genetic studies. The measure for comparing two tests with pool sizes K_1 and K_2 is the Pitman asymptotic relative efficiency (ARE) which is easily seen to

be $\frac{B_{K_2}(h)}{B_{K_1}(h)}$. For the special case of $D = 0$, the ARE is always 1, indicating that increasing the pool size has little impact on the efficiency. In the next section, simulation studies are conducted to compare power of tests for varying linkage disequilibrium coefficients and pool sizes.

Alternatively, we can use the odds ratio $\frac{h_1 h_4}{h_2 h_3}$, or its logarithm $\beta = \log \frac{h_1 h_4}{h_2 h_3}$, to quantify the

linkage disequilibrium. Let $\hat{\beta} = \log \frac{\hat{h}_1 \hat{h}_4}{\hat{h}_2 \hat{h}_3}$ be the maximum likelihood estimator of β . Again, we can apply the asymptotic normality for $\hat{\mathbf{h}}$ and the δ -method to show that $n^{1/2} (\hat{\beta} - \beta)$ converges in distribution to normal with mean 0 and variance

$$A_K(\mathbf{h}) = \frac{(h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})}{2K - v_k(h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})}.$$

For hypothesis testing $H_0 : \beta = \beta_0$ versus $H_a : \beta \neq \beta_0$, we define B test as

$$\frac{\sqrt{n}(\hat{\beta} - \beta_0)}{\sqrt{A_K(\hat{\mathbf{h}})}}$$

Denote $A_K(h)$ by $A_K^1(h)$ in the special case of linkage equilibrium ($\beta = 0$),

$A_K^1(h) = h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}$, which again does not depend on the pool size K . This variance formula is also closely related to the maximum likelihood estimator for the log odds ratio in analysis of a two-by-two table with fixed margins [16,17]. In fact, it is exactly equal to the limiting variance of the latter when the 2×2 table is formed from n independent haplotypes.

5. SIMULATION STUDIES

Monte Carlo studies are conducted to evaluate the accuracy of our variance formula. We consider the case of two loci and took the minor allele frequencies as $p_A = 0.5$, $p_B = 0.4$ and set LD coefficient to be 0, 0.05, or 0.1. For each pool size $K = 1, 2, 3, 4$, we generated $n \times 2K$ haplotypes from the underlying haplotype frequencies and then randomly pooled every $2K$ haplotypes to form pooled genotypes. Based on $n = 100$ pooled genotypes, haplotype frequencies are estimated by EM algorithm. Variances for estimated haplotype frequencies are estimated by the analytical formula in [4] (**AY**) and our formula (**CT**) respectively. The procedure was replicated 1,000 times. Empirical variances of haplotype frequency estimates from 1000 simulations are calculated as (**SV**). The results of these studies for haplotypes (**AB**), (**Ab**) are summarized in Table I. All three estimates are very close. For varying LD coefficients, variances of estimated haplotype frequencies decrease with pool size K and again demonstrated efficiency of pooling in estimating haplotype frequencies.

We also carried out simulation studies to assess performance of LD tests. Allele frequencies are specified the same as above. For various LD coefficients (between 0 and 0.1) and pool sizes $K = 1, 2, 3, 4$, we generate pooled genotypes with sample size $n = 50$ and $n = 200$. For

1,000 replicates, Figure 1 illustrates simulated power of D test and B test (defined in section 4). The power curves are nearly identical for different pool sizes. This can also be seen from theoretical formula of the power. Because $\hat{D}$ and $\hat{\beta}$ are both asymptotically normal, theoretical power of tests can also be calculated from normal distribution. The theoretical

power for the D test is given as $\Phi\left(\frac{\sqrt{n}(D - D_0)}{\sqrt{B_K(h)}} - z_{\alpha/2}\right) + \Phi\left(-\frac{\sqrt{n}(D - D_0)}{\sqrt{B_K(h)}} - z_{\alpha/2}\right)$ for given LD coefficient D and allele frequency p_A and p_B . Theoretical power for the B test is similar. For example for sample size $n = 100$ and $D = 0, 0.05, 0.075, 0.10$, based on the asymptotic normality of the tests, the power for D test is 0.050, 0.174, 0.555, 0.917, 0.998 when pool size $K = 1$. For pool sizes $K = 2, 3, 4$, power is about the same. For example, when $K = 4$, power is 0.050, 0.175, 0.555, 0.918, 0.999. For B test, when $K = 1$, power is 0.050, 0.172, 0.527, 0.870, 0.987. When $K = 4$, power is 0.050, 0.173, 0.527, 0.870, 0.988. The results from normal calculations corroborates that pooling is not efficient in testing linkage disequilibrium.

We also investigate the influence of genotype error on the proposed LD tests. The assay error rates are reported to be in the range of 0.1%–5% with the current individual genotyping technologies [18,19]. We simulated genotypes by allowing normal error, the variances for pooled genotype is assumed to be proportional to the pool size, i.e., we allow larger error for larger pools. The error rate for individual genotype is set to be 0.1%, genotype error rates for $K = 2, 3, 4$ are respectively 0.011, 0.040 and 0.077. Type I error is well controlled to be close to the nominal value of 0.05. Power of the proposed LD tests is illustrated in Figure 2. Compared with the power curve without genotype error in Figure 1, there is minor power loss due to genotype error. The maximal power loss is 8%. For larger pool sizes, power loss is more evident due to larger error in pooled genotypes. Finally, we used simulation studies to examine the sensitivity of tests for linkage disequilibrium to the assumption of Hardy-Weinberg equilibrium. We consider the following form of Hardy-Weinberg disequilibrium:

$$\pi_{kl} = \begin{cases} \pi_k^2 + \rho\pi_k(1 - \pi_k), & k=l \\ (1 - \rho)\pi_k\pi_l, & k \neq l \end{cases}$$

where ρ is the inbreeding coefficient [20]. We set $\rho = 0.05, 0.2$ respectively and simulated 1,000 samples each with 100 pooled genotypes. Simulated power of D test and B test is summarized in Table II. Results show that tests still perform well for slight deviation from the HWE assumption ($\rho = 0.05$). When $\rho = 0.2$, which is unreasonably large for population data, the type I error of the LD tests is still well controlled and both tests are conservative.

6. TWO REAL EXAMPLES

Now we apply our new methods to two real data sets. The first data set is an individual genotype data set ($K = 1$) for angiotensinogen gene (AGT) studied by Jain et al. [21]. The second data set is a pooled AGT gene data set studied by Yang et al. [4].

The proposed Wald-type tests are for general pooled genotype data. As a special case, it can be used for individual genotype data (without phase information, $K = 1$) which is very popular in haplotype based population studies. We first consider the individual AGT genotype data analyzed by Jain et al. [21] for studying the association of AGT gene with hypertension in African-Americans. Hypertension is a serious health problem in western countries, in particular for the African-American population. The AGT gene polymorphism has been shown to play a significant role in hypertension in African-Americans. By comparing allele frequencies in case and control groups at marker SNPs, AGT gene

polymorphism -217 (A/G) was identified to be associated with hypertension in African-Americans. Here we focus on haplotype estimation and linkage disequilibrium tests from individual genotypes without phase information. We estimate haplotype frequencies and conduct linkage disequilibrium tests from genotypes of 46 unrelated individuals (female normal Africans) at two SNPs. The estimated haplotype frequencies, linkage disequilibrium coefficients and tests are summarized in Table III. Both D test and B test suggest that there is no strong linkage disequilibrium between two SNPs.

In the second example, we apply our methods to a real pooled AGT gene data set. Data are the pooled genotypes at 10 SNPs in the AGT gene [4]. Each pool has two individuals. All individuals are unrelated. There are 15 pools and hence 30 individuals, but individual genotypes are not available. We conduct the linkage disequilibrium tests for each pair of four tagging SNPs. The results are summarized in Table IV. Both D test and B test suggest that loci 1 and 4 are in the strongest linkage disequilibrium with test statistics 4.174 and 4.106 (corresponding p -values are respectively 3×10^{-5} and 4×10^{-5}). They are highly significant even after Bonferroni correction is applied to the 6 tests conducted. For these two loci, the estimated frequencies for haplotypes (0, 0), (1, 0), (0, 1), (1, 1) are 0.729, 0.021, 0.060, 0.190 respectively. The strong linkage disequilibrium between them can also be seen from presence of the two small frequencies for haplotypes (1, 0) and (0, 1). On the other hand, as an example, the haplotype frequencies for loci 2 and 3 are estimated to be 0.325, 0.291, 0.156, 0.228, none of which appears to be very small, which, again, is consistent with the insignificance of the test results.

Further examination of Table IV shows loci 1, 3 and 4 are in relatively tight linkage disequilibrium with each others whereas they are all in weak linkage disequilibrium with locus 2. Hence loci 1, 3, and 4 are likely to form a cluster.

7. CONCLUDING REMARKS

We formulate the haplotype estimation problem in terms of a sequence of 2×2 contingency tables with missing cell counts and observed margins. Close connection of the covariance matrix for the maximum likelihood estimate with the extended hypergeometric distribution is built and numerical advantages for variance estimate are obtained. The new perspective also enables us to derive novel linkage disequilibrium tests for pooled genotypes. Thus, efficiency of pooling can be investigated in both estimating haplotype frequencies and testing linkage disequilibrium. It is intriguing to find that pooling is efficient in estimating haplotype frequencies but not efficient in testing weak linkage disequilibrium.

In our studies, efficiency is measured solely by genotyping costs. In other situations where sampling costs need to be considered, pooling is expected to be less efficient and a formal cost-effective analysis can be conducted as in [2]. We have made the Hardy-Weinberg equilibrium assumption which is usual for haplotype inference from population data and we assumed there is no genotype error in the mathematical derivations. Simulation studies show that proposed tests perform reasonably well for slight deviations from these assumptions. In deriving the LD tests, we have assumed that there is no rare haplotypes, i.e., all the four haplotype frequencies are greater than 0. It is very possible this assumption is violated if the loci are in high LD. But for testing for no LD, this assumption is necessary since otherwise one of the two loci must be a non-polymorphism with zero allele frequency, which is not of interest in genetic or statistical analysis. In principle, our approach can be extended to the general m -loci by relating the problem to the generalized 2^m -table with m fixed margins.

APPENDICES

Appendix A: Derivation of $I_K(\mathbf{h})$

Since $p_{\mathbf{h}}(\mathbf{x}) = \sum_{\xi \in H_x} p_{\mathbf{h}}(\xi) = \sum_{\xi \in H_x} \binom{2K}{\xi_1, \xi_2, \dots, \xi_{2^m}} h_1^{\xi_1} h_2^{\xi_2} \dots h_{2^m}^{\xi_{2^m}}$, we have

$$\frac{\partial \log p_{\mathbf{h}}(\mathbf{x})}{\partial \mathbf{h}} = \frac{\sum_{\xi \in H_x} \binom{2K}{\xi_1, \xi_2, \dots, \xi_{2^m}} h_1^{\xi_1} h_2^{\xi_2} \dots h_{2^m}^{\xi_{2^m}} \text{diag} \left\{ \frac{\xi_1}{h_1}, \dots, \frac{\xi_{2^m}}{h_{2^m}} \right\}}{\sum_{\xi \in H_x} p_{\mathbf{h}}(\xi)} = \frac{\sum_{\xi \in H_x} p_{\mathbf{h}}(\xi) \Gamma \xi}{p_{\mathbf{h}}(\mathbf{x})} = E(\Gamma \xi | \mathbf{x})$$

and

$$\begin{aligned} -\frac{\partial^2 \log p_{\mathbf{h}}(\mathbf{x})}{\partial \mathbf{h} \partial \mathbf{h}^T} &= -\frac{\sum_{\xi \in H_x} p_{\mathbf{h}}(\xi) \{\Gamma \xi\}^{\otimes 2} - \Gamma \xi \Gamma}{p_{\mathbf{h}}(\mathbf{x})} + \frac{\left(\sum_{\xi \in H_x} p_{\mathbf{h}}(\xi) \Gamma \xi \right)^{\otimes 2}}{p_{\mathbf{h}}(\mathbf{x})^2} \\ &= \left\{ E(\Gamma \xi | \mathbf{x}) \right\}^{\otimes 2} + \Gamma \text{diag} \left\{ E(\xi | \mathbf{x}) \right\} \Gamma - E \left\{ (\Gamma \xi)^{\otimes 2} | \mathbf{x} \right\} \\ &= \Gamma \text{diag} \left\{ E(\xi | \mathbf{x}) \right\} \Gamma - \text{cov}(\Gamma \xi | \mathbf{x}). \end{aligned} \quad (5)$$

Taking expectation with respect to $\mathbf{x}$ on both sides of (5) and noting that $\Gamma \text{diag} \{E(\xi)\} \Gamma = 2K\Gamma$, we get,

$$I_K(\mathbf{h}) = 2K\Gamma - \Gamma E\{\text{var}(\xi | \mathbf{x})\} \Gamma.$$

Appendix B: Proof of proposition

By its definition, the limiting covariance can be expressed as

$$\sum_K = W \left[2K \left\{ \begin{pmatrix} \frac{1}{h_1} & & \\ & \frac{1}{h_2} & \\ & & \frac{1}{h_3} \end{pmatrix} + \frac{1}{h_4} \mathbf{1}^{\otimes 2} \right\} - v_K \left\{ \begin{pmatrix} \frac{1}{h_1} \\ -\frac{1}{h_2} \\ -\frac{1}{h_3} \end{pmatrix} - \frac{1}{h_4} \mathbf{1} \right\}^{\otimes 2} \right]^{-1} W^T.$$

Define 3×3 matrix $A = \text{diag}(h_1^{-1}, h_2^{-1}, h_3^{-1}) + h_4^{-1} \mathbf{1}^{\otimes 2}$ and 3×1 vector $\mathbf{q} = (-h_1^{-1}, h_2^{-1}, h_3^{-1})^T + h_4^{-1} \mathbf{1}$. We thus have

$$\begin{aligned} \sum_K &= W(2KA - v_K \mathbf{q}^{\otimes 2})^{-1} W^T \\ &= \frac{1}{2K} W \left(A^{-1} + \frac{v_K}{2K} \frac{A^{-1} \mathbf{q}^{\otimes 2} A^{-1}}{1 - \frac{v_K}{2K} \mathbf{q}^T A^{-1} \mathbf{q}} \right) W^T. \end{aligned} \quad (6)$$

It is easy to verify that $q^T A^{-1} = (1, -1, -1)$ and $q^T A^{-1} q = h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}$. Therefore,

$$\begin{aligned}
 \sum_K &= \frac{1}{2K} W \left\{ A^{-1} + \frac{\frac{v_K}{2K}}{1 - \frac{v_K}{2K} \left(\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right)} \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}^{\otimes 2} \right\} W^T \\
 &= \frac{1}{2K} \left\{ \Gamma^{-1} - \mathbf{h}^{\otimes 2} + \frac{\frac{v_K}{2K}}{1 - \frac{v_K}{2K} \left(\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right)} \mathbf{u}_1^{\otimes 2} \right\} \\
 &= \frac{1}{2K} \left[\sum_{\frac{1}{2}} + \frac{\frac{v_K}{2K}}{1 - \frac{v_K}{2K} \left(\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right)} \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \right].
 \end{aligned} \tag{7}$$

Appendix C: Proof of Corollary 1

From the well-known variance formula for the (central) hypergeometric distribution it follows that, when $\psi = 1$,

$$\text{var}(\xi_1 | \mathbf{x}) = \frac{x_1 x_2 y_1 y_2}{(2K)^2 (2K - 1)}.$$

Since $(\xi_1, \xi_2, \xi_3, \xi_4) \sim M(2K, (h_1, h_2, h_3, h_4))$ and $E(x_1 x_2 y_1 y_2) = E\{(\xi_1 + \xi_2)(\xi_3 + \xi_4)(\xi_1 + \xi_3)(\xi_2 + \xi_4)\}$, we can show that

$$v_K^1 = E\{\text{var}(\xi_1 | \mathbf{x})\} = (2K - 1) h_1 h_4 = \frac{2K - 1}{h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}}. \tag{8}$$

Combining (8) with the Proposition yields

$$\begin{aligned}
 \sum_K^1 &= \frac{1}{2K} \sum_{\frac{1}{2}} + \frac{2K-1}{2K} \frac{1}{h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}} \mathbf{u}_1^{\otimes 2} \\
 &= \frac{1}{2K} \left(\sum_{\frac{1}{2}} - \frac{1}{h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}} \mathbf{u}_1^{\otimes 2} \right) + \frac{1}{h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1}} \mathbf{u}_1^{\otimes 2}.
 \end{aligned} \tag{9}$$

It is easy to verify that $\sum_{\frac{1}{2}} - (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})^{-1} \mathbf{u}_1^{\otimes 2} \geq 0$. Thus $\sum_K^1$ decreases as K

increases and, in particular, $\sum_K^1 \leq \sum_{\frac{1}{2}}$. Moreover, as

$$K \rightarrow \infty, \sum_K^1 \rightarrow (h_1^{-1} + h_2^{-1} + h_3^{-1} + h_4^{-1})^{-1} \mathbf{u}_1^{\otimes 2}.$$

Acknowledgments

This research was supported in part by grants from the U.S. National Institutes of Health, the U.S. National Science Foundation and the China Natural Science Foundation (No. 10671189, No. 30730057) and the Knowledge Innovation Program of the Chinese Academy of Science.

REFERENCES

1. Norton N, Williams NM, O'Donovan MC, Owen MJ. DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.* 2004; 36:146–152. [PubMed: 15119834]

2. Wang S, Kidd KK, Zhao H. On the use of DNA pooling to estimate haplotype frequencies. *Genet. Epidemiol.* 2003; 24:74–82. [PubMed: 12508258]
3. Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, Kamatani N. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am. J. Hum. Genet.* 2003; 72:384–398. [PubMed: 12533787]
4. Yang Y, Zhang J, Hoh J, Matsuda F, Xu P, Lathrop M, Ott J. Efficiency of SNP haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA.* 2003; 100:7225–7230. [PubMed: 12777616]
5. Xie X, Ott J. Testing linkage disequilibrium between a disease and marker loci. *Am. J. Hum. Genet.* 1993; 53:1170.
6. Excoffier L, Slatkin M. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 1995; 12:921–927. [PubMed: 7476138]
7. Harkness WL. Properties of the extended hypergeometric distribution. *Ann. Math. Statist.* 1965; 36:938–945.
8. Liao JG, Rosen O. Fast and Stable Algorithms for Computing and Sampling From the Noncentral Hypergeometric Distribution. *Amer. Stat.* 2001; 55:366–369.
9. Borkowf CB. Computing the non-null asymptotic variance of n -way contingency tables with fixed one-way marginal totals, with application to gene-environment interaction studies. *Statist. Sinica.* 2002; 12:491–504.
10. Lehmann, EL. *Elements of Large-Sample Theory.* New York: Springer-Verlag; 2004.
11. Lange, K. *Numerical Analysis for Statisticians.* Berlin: Springer-Verlag; 1999.
12. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian S, Ward R, Lander ES. Linkage disequilibrium in the human genome. *Nature.* 2001; 411:199–204. [PubMed: 11346797]
13. Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 2005; 6:109–118. [PubMed: 15716907]
14. Pfeiffer RM, Rutter JL, Gail MH, Struwing J, Gastwirth JL. Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.* 2002; 22:94–102. [PubMed: 11754476]
15. Serfling, RJ. *Approximation Theorems of Mathematical Statistics.* New York: Wiley; 1980.
16. Breslow, NE.; Day, NE. *Statistical Methods in Cancer Research, Vol. I, The Design and Analysis of Case-Control Studies.* Lyon: IARC; 1980.
17. Kou SG, Ying Z. Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica.* 1996; 6:809–829.
18. Hao K, Li C, Rosenow C, Wong WH. Estimation of genotype error rate using samples with pedigree informationCan application on the GeneChip Mapping 10K array. *Genomics.* 2004; 84:623–630. [PubMed: 15475239]
19. Tintle NL, Ahn K, Mendell NR, Gordon D, Finch SJ. Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and Center for Inherited Disease Research. *BMC Genet.* 2005; 6 Suppl 1:S154. [PubMed: 16451615]
20. Zeng D, Lin D. Estimating Haplotype-Disease Associations With Pooled Genotype Data. *Genet. Epidemiol.* 2005; 28:70–82. [PubMed: 15558554]
21. Jain S, Tang X, Narayanan C, Agarwal Y, Peterson SM, Brown CD, Ott J, Kumar A. Angiotensinogen gene polymorphism at -217 affects basal promoter activity and is associated with hypertension in African-Americans. *J. Biol. Chem.* 2002; 277:36889–36896. [PubMed: 12145290]

Figure 1.
Simulated power of LD tests.

Figure 2.
Simulated power of LD tests when genotype error is present.

Table I

Simulation results for variance estimates

Haplotype	Allele frequency	D	K	SV	AY	CT
11	0.2	0	1	0.118	0.110	0.110
			2	0.092	0.084	0.085
			4	0.072	0.074	0.072
	0.25	0.05	1	0.119	0.119	0.118
			2	0.081	0.085	0.086
			4	0.075	0.069	0.070
	0.3	0.1	1	0.120	0.118	0.118
			2	0.079	0.076	0.076
			4	0.055	0.057	0.056
10	0.2	0.1	1	0.101	0.094	0.093
			2	0.068	0.064	0.064
			4	0.048	0.050	0.050
	0.25	0.05	1	0.115	0.119	0.118
			2	0.086	0.086	0.086
			4	0.073	0.069	0.070
	0.3	0	1	0.134	0.134	0.135
			2	0.104	0.098	0.098
			4	0.086	0.080	0.079

Table II
Simulated power of LD tests under deviations of HWE at significance level 0.05

ρ	Test	D				
		K	0.000	0.025	0.050	0.075 0.100
0	D test	1	0.053	0.191	0.540	0.897 0.992
		2	0.054	0.182	0.583	0.900 0.996
		3	0.056	0.186	0.558	0.904 0.991
		4	0.049	0.191	0.533	0.901 0.992
	B test	1	0.056	0.173	0.508	0.888 0.990
		2	0.046	0.167	0.564	0.893 0.995
		3	0.057	0.168	0.539	0.897 0.991
		4	0.055	0.172	0.504	0.891 0.988
	D test	1	0.066	0.193	0.541	0.916 0.989
		2	0.063	0.176	0.527	0.874 0.987
		3	0.062	0.166	0.512	0.909 0.989
		4	0.053	0.170	0.536	0.899 0.993
0.05	B test	1	0.060	0.178	0.520	0.906 0.988
		2	0.057	0.161	0.512	0.867 0.986
		3	0.055	0.148	0.487	0.897 0.987
		4	0.046	0.154	0.508	0.889 0.990
	D test	1	0.078	0.208	0.542	0.892 0.990
		2	0.029	0.154	0.465	0.838 0.984
		3	0.034	0.124	0.450	0.828 0.984
		4	0.043	0.123	0.449	0.839 0.975
	B test	1	0.069	0.192	0.522	0.882 0.987
		2	0.024	0.144	0.443	0.820 0.984
		3	0.026	0.108	0.431	0.806 0.983
		4	0.036	0.112	0.425	0.826 0.974

Table III

Haplotype estimation and LD tests for individual AGT data

Haplotype frequency	(1,1) 0.029	(1,0) 0.079	(0,1) 0.134	(0,0) 0.758
D'	D test (p value)		95% C.I.	
0.012	0.596 (0.551)		(-0.024,0.048)	
β	B test (p value)		95% C.I.	
0.750	0.684 (0.494)		(-1.209,2.709)	

Table IV

Linkage disequilibrium tests for pooled AGT data

Test ^a	Locus	Locus			
		1	2	3	4
<i>D</i> test	1	-	0.386 (0.021)	1.331 (0.069)	4.174 (0.138)
	2	-	-	0.423 (0.029)	0.548 (0.032)
	3	-	-	-	1.224 (0.069)
	4	-	-	-	-
<i>B</i> test	1	-	0.375 (0.522)	1.171 (1.748)	4.106 (4.705)
	2	-	-	0.423 (0.487)	0.525 (0.695)
	3	-	-	-	1.132 (1.546)
	4	-	-	-	-

^aEstimated LD coefficient (*D*) and log-odds ratio (β) are in parentheses.