BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 20 2010, pages 2556–2563
doi:10.1093/bioinformatics/btq492

Genetics and population analysis

Advance Access publication August 27, 2010

A study of the efficiency of pooling in haplotype estimation

Anthony Y. C. Kuk[1,*], Jinfeng Xu[1] and Yaning Yang[2]

[1]Department of Statistics and Applied Probability, National University of Singapore,117546 Singapore and
[2]Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: It has been claimed in the literature that pooling DNA samples is efficient in estimating haplotype frequencies. There is, however, no theoretical justification based on calculation of statistical efficiency. In fact, the limited evidence given so far is based on simulation studies with small numbers of loci. With rapid advance in technology, it is of interest to see if pooling is still efficient when the number of loci increases.

Methods: Instead of resorting to simulation studies, we make use of asymptotic statistical theory to perform exact calculation of the efficiency of pooling relative to no pooling in the estimation of haplotype frequencies. As an intermediate step, we use the log-linear formulation of the haplotype probabilities and derive the asymptotic variance–covariance matrix of the maximum likelihood estimators of the canonical parameters of the log-linear model.

Results: Based on our calculations under linkage equilibrium, pooling can suffer huge loss in efficiency relative to no pooling when there are more than three independent loci and the alleles are not rare. Pooling works better for rare alleles. In particular, if all the minor allele frequencies are 0.05, pooling maintains an advantage over no pooling until the number of independent loci reaches 6. High linkage disequilibrium effectively reduces the number of independent loci by ruling out certain haplotypes from occurring. Similar calculations of efficiency for the case of no pooling justify the common belief that it is not worthwhile to use molecular methods to resolve the phase ambiguity of individual genotype data.

Availability: The R codes for the calculation are available at http://www.stat.nus.edu.sg/~staxj/pooling

Contact: stakuka@nus.edu.sg

Received on June 16, 2010; revised on July 29, 2010; accepted on August 23, 2010

1 INTRODUCTION

Pooling DNA samples is a common cost-effective practice in genetic association studies, especially for the purpose of initial screening (Bansal *et al*., 2002; Barcellos *et al*., 1997; Norton *et al*., 2004; Pearson *et al*., 2007; Sham *et al*., 2002; Zuo *et al*., 2006). Large pools of DNA samples also come up in forensic studies (Homer *et al*., 2008; Jacobs *et al*., 2009). To facilitate haplotype-based association analysis, it is necessary to estimate haplotype frequencies from pooled DNA data, and many algorithms, mainly the expectation maximization (EM) algorithm and its variants (Ito *et al*., 2003; Kierkpatrick *et al*., 2007; Kuk *et al*., 2009; Niu, 2004; Wang *et al*., 2003; Yang *et al*., 2003; Zhang *et al*., 2008), have been proposed in the literature. However, the emphasis of these papers is very much algorithmic and numerical, focusing on computing speed and numerical convergence, rather than on the statistical property and efficiency of the estimates being computed. This is unfortunate, because even if the best available algorithm is used to compute, say, the maximum likelihood estimate (MLE), it will be of no use if the estimator itself does not possess good statistical property. What we meant by the statistical property of an estimator is its performance under repeated sampling from the postulated model. Thus, we are concerned about quantities such as bias, variance and also coverage in the case of interval estimation. To compute the estimation efficiency of one estimator relative to another, we adopt the usual definition of taking the ratio of the variance (or mean squared error for biased estimators) of estimator 2 to that of estimator 1. In this article, estimator 1 is the MLE based on pooled genotype data, and estimator 2 is the MLE based on individual genotype data. The number of pools used in estimator 1 and the number of individuals used in estimator 2 are kept equal so that the genotyping costs of the two estimators are the same.

Given the large body of work on haplotype estimation based on pooled data, it will be important for practitioners to know when to pool and when not to pool. Yang *et al.* (2003) believed that pooling DNA samples is efficient in estimating haplotype frequencies. In particular, they demonstrated that although pooling K individuals increases ambiguities, the uncertainty of ML estimation increases less than K times that of unpooled DNA, at least for small pool sizes and small number of loci. Thus, for the same genotyping cost, the pooled data MLE will be more efficient than the MLE computed from individual genotype data. Their comparison was limited to the case of small number of loci. Due to rapid advance in technology, huge numbers of SNPs are genotyped routinely, and it would be interesting to find out what happens when the number of loci increases.

Our study differs from Yang *et al*. (2003) in that our reported efficiencies are based on theoretical calculations using asymptotic variance formulae rather than based on simulations. By carrying out theoretical calculations, we avoid some of the shortcomings of simulation studies and can handle more loci. Our findings, for the case of linkage equilibrium and non-rare allele, suggest that pooling begins to lose estimation efficiency (relative to no pooling at the same genotyping cost) when the number of loci is larger than 3. Other factors affecting the efficiency of pooling that have been mentioned in the literature include sparsity (Barratt *et al*., 2002), linkage disequilibrium (LD) and allele frequencies (Kirk and Cardon, 2002). Barratt *et al*. (2002) commented that it is not necessarily the case that pooling will lead to loss of haplotype

*To whom correspondence should be addressed.

information when LD is so strong that only a small number of haplotypes will occur with appreciable frequency. Kirk and Cardon (2002) defined high haplotype complexity by 'low LD, common alleles' and low haplotype complexity by 'high LD and/or many rare alleles'. Another aim of this article is to unify and reconcile all these claims with our findings. One way to think about LD is that it reduces the 'effective number of independent loci' by reducing the number of non-rare haplotypes. Thus, if the L loci can be grouped into B blocks with high LD within blocks and low LD between blocks, then the situation is similar to the case of B independent loci. We report some efficiency calculations to support this conjecture. The effect of rare alleles is also to reduce the number of non-rare haplotypes to just a few. Instead of having 8, 16 and 32 possible haplotypes for the cases of 3, 4 and 5 loci, if all the minor allele frequencies are small, we will argue later that there are only 4, 5 and 6 haplotypes with non-negligible probabilities when the number of loci is 3, 4 and 5. As a result, it will take a larger number of independent loci for pooling to lose efficiency relative to no pooling for rare alleles. For example, we show that if all the minor allele frequencies are 0.05, then pooling begins to lose efficiency when there are 6 (rather than 4) independent loci. When there is no rare allele, the normal guideline of not to pool when the effective number of independent loci is more than 3 is applicable.

2 THEORY

In order to derive the asymptotic variances of estimated haplotype frequencies, we reparameterize haplotype frequencies by using a log-linear model representation. In this section, we first calculate the asymptotic variances of estimated log odds ratios in the log-linear model under the assumption of linkage equilibrium, then compute the asymptotic variances of estimated haplotype frequencies by using the delta method. Let L be the number of loci in each DNA strand, Y_j the binary allele at locus j, and $\mathbf{Y}=(Y_1,\ldots,Y_L)$ denote the haplotype. There are various ways to parameterize the distribution of multivariate binary data such as $\mathbf{Y}=(Y_1,\ldots,Y_L)$. We shall start with the canonical parameters because they are the simplest mathematically and they are not constrained. Under this log-linear formulation (Fitzmaurice *et al.*, 1993; Liang *et al.*, 1992), the probability distribution of $\mathbf{Y}=(Y_1,\ldots,Y_L)$ is given by

$$P(\mathbf{Y}=\mathbf{y})=c(\alpha)\exp\left(\sum_{j=1}^{L}\alpha_j^{(L)}y_j+\sum_{j_1<j_2}\alpha_{j_1j_2}^{(L)}y_{j_1}y_{j_2}+\sum_{j_1<j_2<j_3}\alpha_{j_1j_2j_3}^{(L)}y_{j_1}y_{j_2}y_{j_3}+\cdots+\alpha_{12\cdots L}^{(L)}y_1y_2\ldots y_L\right), \tag{1}$$

where $\mathbf{y}=(y_1,y_2,\ldots,y_L)$ is a realization of $\mathbf{Y}=(Y_1,\ldots,Y_L)$, $c(\alpha)$ is the normalizing constant and $\Psi=\{0,1\}^L$ is the collection of all possible L-tuples of 0's or 1's. As pointed out by Liang *et al.* (1992), the canonical parameters α have interpretations as the log-conditional odds, log-conditional odds ratios and higher order log-conditional odds ratios, which are defined as contrasts of log-conditional odds ratios. For example,

$$\exp\left(\alpha_1^{(L)}\right)=\frac{P(Y_1=1|Y_2=\cdots=Y_L=0)}{P(Y_1=0|Y_2=\cdots=Y_L=0)},$$

$$\exp\left(\alpha_{12}^{(L)}\right)=\frac{P(Y_1=1,Y_2=1|Y_3=\cdots=Y_L=0)}{P(Y_1=1,Y_2=0|Y_3=\cdots=Y_L=0)}\times\frac{P(Y_1=0,Y_2=0|Y_3=\cdots=Y_L=0)}{P(Y_1=0,Y_2=1|Y_3=\cdots=Y_L=0)},$$

and so on. Note that $\alpha_{12}^{(L)}$ and $\alpha_{12}^{(L-1)}$ have different interpretations because the conditioning sets are different. While the canonical parameters are easier to handle mathematically, it is more meaningful to talk about the unconditional odds and odds ratio

$$\exp(\psi_1)=\frac{P(Y_1=1)}{P(Y_1=0)},$$

$$\exp(\psi_{12})=\frac{P(Y_1=1,Y_2=1)P(Y_1=0,Y_2=0)}{P(Y_1=1,Y_2=0)P(Y_1=0,Y_2=1)},$$

and so on. Because these are parameters for the unconditional marginal distributions, there is no need to add superscripts to the ψ's like what we did for the α's. One useful observation is

$$\alpha_{12\cdots L}^{(L)}=\psi_{12\cdots L}$$

because the conditioning set of $\alpha_{12\cdots L}^{(L)}$ is empty, and so $\alpha_{12\cdots L}^{(L)}$ can be interpreted unconditionally.

By making use of the method of efficient score, we prove in Appendix A (Section A1) that as the number of pools n_P increases, the asymptotic variance of the MLE of $\alpha_{12\cdots L}^{(L)}=\psi_{12\cdots L}$ based on pooled allele frequencies at L loci which are at linkage equilibrium and K individuals in each pool, is given by

$$\frac{(2K)^{L-2}}{n_P\prod_{j=1}^{L}p_j(1-p_j)}, \tag{2}$$

where $p_j=P(Y_j=1)$.

By applying the same argument to $m<L$, we can conclude that the asymptotic variance under linkage equilibrium of the MLE of $\alpha_{12\cdots m}^{(m)}=\psi_{12\cdots m}$ based on pooled data for the first m loci only is also given by (2), with m in place of L. But since the pooled allele frequencies at the remaining $L-m$ loci contain no additional information about $\psi_{12\cdots m}$ under linkage equilibrium, the MLE of $\psi_{12\cdots m}$ based on pooled data at all L loci has the same asymptotic variance as the MLE based on the first m loci alone. This heuristic argument can be made precise and a rigorous proof can be provided on request. Thus the asymptotic variance (under linkage equilibrium) of the MLE of $\psi_{12\cdots m}$ based on the pooled allele frequencies at all L loci will still be given by (2) with m in place of L.

The results presented thus far pertain to the estimation of the unconditional log odds ratios ψ's. What the scientists are more interested in estimating are the haplotype probabilities given by (1). To obtain the asymptotic variance of the haplotype frequency estimates, we first obtain the asymptotic variance–covariance matrix $V_K(\alpha)$ of the pooled data MLE of the canonical parameters α's. An outline of how to do this is given in Appendix A (Section A2). Next, since $P(\mathbf{Y}=y)$ for each y is a function of the canonical parameters according to (1), the asymptotic variance $V_K(y)$ of its MLE is $D_yV_K(\alpha)D_y^T$ by the delta method, where D_y is the derivative of $P(Y=y)$ with respect to α. For example, when $L=2$, $p_{11}=\frac{\exp(\alpha_1+\alpha_2+\alpha_{12})}{1+\exp(\alpha_1)+\exp(\alpha_2)+\exp(\alpha_1+\alpha_2+\alpha_{12})}$, $\frac{\partial p_{11}}{\partial\alpha_1}=p_1p_2(1-p_1)$, $\frac{\partial p_{11}}{\partial\alpha_2}=p_1p_2(1-p_2)$, $\frac{\partial p_{11}}{\partial\alpha_{12}}=p_1p_2(1-p_1p_2)$, and

the asymptotic variance of the MLE of p_{11} is $\frac{1}{n_P}\{\frac{p_1p_2(1-p_1p_2)}{2K}+\frac{2K-1}{2K}p_1p_2(1-p_1)(1-p_2)\}$.

3 RESULTS

3.1 Efficiency in estimating unconditional log odds ratios

We first compute the asymptotic relative efficiency (ARE) of the pooled data MLE of $\psi_{12\cdots m}, m \le L$, to the unpooled data MLE (i.e. $K=1$) as the following ratio of two asymptotic variances

$$\mathrm{ARE}(\widehat{\psi}_{12\cdots \mathrm{m}}) = \left\{ \frac{(2)^{\mathrm{m}-2}}{\mathrm{n_P}\prod_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{p_j}(1-\mathrm{p_j})} \right\} \Big/ \left\{ \frac{(2\mathrm{K})^{\mathrm{m}-2}}{\mathrm{n_P}\prod_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{p_j}(1-\mathrm{p_j})} \right\} = K^{2-m}. \quad (3)$$

It follows that for estimating the locus-specific log odds ψ_j, corresponding to $m=1$, ARE is given by $K^{2-1}=K$ which is always greater than 1 since the pool size K is at least 2. Thus, pooling gains efficiency for estimating first-order quantities like ψ_j, bearing in mind that we are not comparing the estimate based on n pools of K individuals each with the estimate based on nK individuals, but rather with the estimate based on n individuals, so that the genotyping costs are the same.

For estimating the unconditional log odds ratio ψ_{jk} at two loci, $m=2$, and the ARE is $K^{2-2}=1$, meaning that the pooled and unpooled data MLEs are equally efficient. For estimating ψ_{jkl}, the ARE is $K^{2-3}=K^{-1}<1$, and so pooling will lose efficiency. For estimating ψ_{jklm}, the ARE is $K^{2-4}=K^{-2}$, which is worse.

While it is not a big surprise that we cannot estimate higher order association parameters well from pooled DNA data, it is quite amazing to discover that there is an efficiency ladder. It is neat to be able to work out precisely the exact orders (in powers of K, the pool size) of the asymptotic variance of the pooled data MLE of the different parameters. Equation (3) is revealing because it tells us that the ARE of pooling versus no pooling in estimating the $(m-1)$-th order log odds-ratio $\psi_{12\ldots m}$ is reduced by a factor of K whenever m is increased by 1.

Table 1 summarizes what happen for up to the case of 8 loci. The numbers in the first column are the numbers of ψ parameters for which pooling gains efficiency in estimation. In the next column are the numbers of ψ parameters for which pooling and no pooling are equally efficient. Further right are the numbers of ψ parameters for which pooling loses efficiency in estimation, and notice that the farther out we go to the right of the table, the heavier is the loss in efficiency. We can see from Table 1 that for the case of two loci, there are no ψ parameters with ARE less than 1. For the case of three loci, there are three ψ_j's for which the pooled data MLE is better, three ψ_{jk}'s for which pooling and no pooling are asymptotically equivalent, and pooling is worse (ARE$=K^{-1}$) only for the estimation of ψ_{123}. These explain the findings of Yang *et al.* (2003) that pooling is efficient in estimating the haplotype probabilities $P(\mathbf{Y}=y)$, which are functions of the ψ's, when there are only two or three loci. But Table 1 also tells us that the same cannot be expected to hold true when the number of loci is 4 or more.

Table 1. Grouping the ψ parameters according to the ARE with which they can be estimated by the pooled data MLE for the case of $L=2$ to 8 loci

ARE	K	1	K^{-1}	K^{-2}	K^{-3}	K^{-4}	K^{-5}	K^{-6}
L=2	2	1						
L=3	3	3	1					
L=4	4	6	4	1				
L=5	5	10	10	5	1			
L=6	6	15	20	15	6	1		
L=7	7	21	35	35	21	7	1	
L=8	8	28	56	70	56	28	8	1

The numbers shown are the number of parameters in each group and K is the pool size.

For the case of four loci, there are four ψ parameters which can be estimated from pooled data with ARE=$K>1$, 6 parameters with ARE$=1$, 4 parameters with ARE$=\mathrm{K}^{-1}<1$, and 1 parameter with ARE$=\mathrm{K}^{-2}<<1$. On this ground, we expect pooling to begin to lose efficiency in the estimation of haplotype frequencies when there are four loci. This is confirmed by theoretical calculations to be reported later. Things should get progressively worse as the number of loci L increases. One can see from Table 1 that as L increases, there will be more and more higher order parameters for which pooling will do progressively worse, since the ARE decreases by a factor of K every time we move to the right by one column.

3.2 Efficiency for haplotype frequency estimates

For haplotype frequency estimation, we can define the asymptotic efficiency of pooling relative to no pooling by

$$\sum_{y\in\Psi} V_1(y) \Big/ \sum_{y\in\Psi} V_K(y), \quad (4)$$

where K is the number of individuals in each pool, $V_K(y)$ is the variance of the MLE of $P(Y=y)$, and $\Psi=\{0,1\}^L$. As described at the end of Section 2, this ratio can be calculated theoretically for different choices of the number of loci L, minor allele frequencies and pool size K. To reduce the number of configurations, we select the minor allele frequencies to be equally spaced between 0.1 and 0.3. The results are summarized in Table 2. It can be seen that pooling loses efficiency in the estimation of haplotype frequencies when the number of loci is 4 or more, as one could have guessed from Table 1. Pooling will lose efficiency even for the case of three loci, when the pool size reaches 5 or more. For every fixed pool size K (i.e. column of Table 2), the efficiency of pooling decreases rapidly as the number of loci L increases. It decreases with pool size when $L\ge 3$. The efficiency loss can be really huge. In fact, the efficiency of pooling relative to no pooling is no greater than 20% when the number of loci is 5 or more and the pool size is at least 4.

To show how the variances of the haplotype frequency estimates grow as the number of loci increases, the lower panel of Table 2 displays the sum of the asymptotic variances of the haplotype frequency estimates. For calibration purposes, the figures are divided by 2^L (the total number of possible haplotypes) so that they can be interpreted as the average variances over all possible haplotypes. The numbers are further multiplied by n_P, the number of pools, to remove their dependence on n_P. We observe that the average

Table 2. Pooling versus no pooling in estimating haplotype frequencies for various combinations of L (number of loci) and K (pool size) under linkage equilibrium and equally spaced minor allele frequencies

L	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
ARE								
2	1	1.597	1.995	2.278	2.490	2.655	2.787	2.894
3	1	1.162	1.134	1.059	0.978	0.902	0.834	0.773
4	1	0.826	0.620	0.469	0.364	0.289	0.234	0.194
5	1	0.577	0.330	0.201	0.131	0.089	0.064	0.047
6	1	0.398	0.173	0.085	0.046	0.027	0.017	0.011
7	1	0.273	0.090	0.036	0.016	0.008	0.005	0.003
8	1	0.186	0.047	0.015	0.006	0.002	0.001	0.001
Average variance multiplied by n_P								
2	0.075	0.047	0.038	0.033	0.030	0.028	0.027	0.026
3	0.059	0.051	0.052	0.056	0.060	0.065	0.071	0.076
4	0.042	0.051	0.068	0.090	0.115	0.145	0.179	0.217
5	0.029	0.050	0.087	0.142	0.219	0.320	0.450	0.611
6	0.019	0.048	0.111	0.225	0.414	0.704	1.127	1.718
7	0.013	0.047	0.141	0.356	0.781	1.544	2.818	4.823
8	0.008	0.045	0.180	0.562	1.472	3.385	7.041	13.53

Table 3. Relative efficiency of pooling versus no pooling under linkage equilibrium for the case with all minor allele frequencies equal to P

L	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
$P=0.5$								
2	1	1.333	1.500	1.600	1.667	1.714	1.750	1.778
3	1	0.839	0.684	0.571	0.489	0.426	0.378	0.339
4	1	0.513	0.300	0.195	0.137	0.101	0.077	0.061
5	1	0.310	0.130	0.066	0.038	0.069	0.052	0.040
6	1	0.186	0.090	0.042	0.023	0.071	0.053	0.041
7	1	0.140	0.086	0.039	0.021	0.090	0.067	0.052
8	1	0.126	0.097	0.043	0.023	0.128	0.095	0.073
$P=0.1$								
2	1	1.695	2.206	2.597	2.907	3.158	3.365	3.540
3	1	1.408	1.558	1.589	1.565	1.517	1.457	1.394
4	1	1.150	1.064	0.928	0.796	0.683	0.587	0.509
5	1	0.925	0.707	0.589	0.463	0.371	0.302	0.250
6	1	0.744	0.479	0.416	0.311	0.240	0.190	0.154
7	1	0.609	0.341	0.324	0.219	0.002	0.137	0.110
8	1	0.511	0.257	0.271	0.192	0.143	0.110	0.087
$P=0.05$								
2	1	1.826	2.521	3.113	3.623	4.068	4.459	4.805
3	1	1.656	2.085	2.361	2.533	2.635	2.688	2.706
4	1	1.492	1.699	1.749	1.716	1.641	1.548	1.447
5	1	1.336	1.379	1.297	1.176	1.049	0.930	0.825
6	1	1.189	1.145	0.981	0.835	0.708	0.602	0.515
7	1	1.056	0.935	0.002	0.619	0.506	0.001	0.350
8	1	0.936	0.789	0.150	0.479	0.381	0.308	0.254

variance decreases with L when the pool size is 1. They are more or less constant when the pool size is 2. For pool size larger than 2, the average variance increases with L, and more rapidly the larger L is. Note that if we divide the first column of the lower panel of Table 2 by the remaining columns, we recover the relative efficiencies displayed in the upper panel which always decay when the number of loci increases regardless of pool size.

In addition to the scenario of equally spaced minor allele frequencies between 0.1 and 0.3, we also follow Kirk and Cardon (2002) and consider the cases where the minor allele frequencies are 0.5, 0.1 or 0.05 across all markers. The first situation will lead to many distinct haplotypes and the last situation will lead to very few haplotypes with appreciable probabilities and so they are at two ends of the sparsity spectrum. Based on our earlier argument, we expect pooling to fare better for rare alleles and this is confirmed by Table 3. Note that there is severe loss in efficiency due to pooling, even for the case of three loci, when the minor allele frequency is 0.5. For allele frequency 0.1, pooling begins to lose efficiency relative to no pooling when the number of loci is 4 or more. However, when the minor allele frequency is 0.05, we observe from Table 3 that pooling is more efficient than no pooling until we reach six or seven loci, and not the usual four loci that we have been advocating so far.

There is an explanation why pooling is expected to fare better for rare alleles. Let p be the minor allele frequency which is assumed to be small and constant over loci, the 2^L possible haplotypes can be grouped according to the size of their probabilities of occurrence. Obviously, $P(0,\ldots,0)=(1-p)^L$ is of order $O(1)$ as p goes to zero. The next frequent batch of haplotypes are the unit vectors $(1,0,\ldots,0),\ldots,(0,\ldots,0,1)$, containing one '1' and $L-1$ zeros with probability of occurrence $P(1,0,\ldots,0)=p(1-p)^{L-1}=O(p)$. All other haplotype probabilities will be of order $O(p^2)$ or smaller. Thus, out of the possible 2^L haplotypes, only $L+1$ of them have non-negligible probabilities of occurrence. Following through the above argument, we have $P(1,0,\ldots,0)\approx P_1(1),\ldots,P(0,\ldots,0,1)\approx P_L(1)$, where $P_1(1),\ldots,P_L(1)$ are the locus-specific probability for allele '1'. These locus-specific or marginal probabilities can be estimated accurately from the corresponding sample proportions and the precision of a sample proportion depends only on the number of individuals in the sample whether it is pooled or not. In our setup of equal genotyping cost, there are K times more individuals under pooling compared with no pooling, where K is the pool size. Thus, it is not surprising that pooling does better than no pooling in estimating haplotype frequencies for very rare alleles. As to how rare is rare, Table 3 suggests that a minor allele frequency of 0.1 is still within the domain of applicability of the usual guideline. When all the minor allele probabilities are 0.05, pooling has an advantage over no pooling until we reach six or seven loci. When half the minor allele frequencies are 0.01 and the other half are 0.05 (not reported in Table 3), pooling only begins to lose efficiency when the number of loci is 10. Further calculations suggest that pooling is more efficient than no pooling when all the minor allele frequencies equal 0.01.

3.3 Effects of genotyping error

We have assumed perfect genotype data thus far. In practice, there will be error in genotyping. The effects of not accounting for genotyping errors in haplotype estimation have been studied by a number of authors using different models of genotyping errors, which are assumed to occur independently across markers. Kirk and Cardon (2002) reported that 'genotyping error can significantly decrease haplotype frequency and reconstruction accuracy'. The emphasis of that paper was on the comparison of genotyping families versus unrelated individuals in the presence of genotyping error

Table 4. Relative efficiency for the case of equally spaced minor allele frequencies with genotyping error $\epsilon_0=\epsilon_1=0.05$ and $n_P=50$

L	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
2	1	1.336	1.504	1.606	1.673	1.721	1.758	1.786
3	1	1.113	1.094	1.042	0.984	0.927	0.873	0.824
4	1	0.858	0.675	0.530	0.421	0.341	0.281	0.234
5	1	0.619	0.370	0.231	0.152	0.105	0.075	0.055
6	1	0.427	0.191	0.095	0.052	0.031	0.019	0.013
7	1	0.289	0.097	0.039	0.018	0.009	0.005	0.003
8	1	0.194	0.049	0.016	0.006	0.003	0.001	0.001

and pooling was not considered. By using a uniform distribution of error for 0–10% randomly chosen individuals for whom Pr(observed genotype | underlying genotype) is the same for all possible genotypes, they might have exaggerated the effect of genotyping error by giving too much weight to large errors. More relevant to our theme of comparing pooling with no pooling, Quade *et al.* (2005) considered estimation of haplotype frequencies from pooled DNA samples when there is genotyping error and concluded that 'the EM algorithm performs well even in the presence of genotyping error' and that genotyping error only 'slightly decreases the accuracy of haplotype frequency estimates'. However, they studied the two-marker case only and their αn-binomial model of observed genotype given true genotype was primarily an inflated variance model rather than one inducing a systematic bias. A third model of genotyping error proposed by Zou and Zhao (2003) assumed that genotyping error was introduced independently into each marker of each chromosome. We will use this model to demonstrate the effect of genotyping error because it is more amenable to theoretical analysis. For biallelic (0 and 1) marker, let ϵ_0 be the probability of miscalling a '0' as a '1', and ϵ_1 the probability of miscalling a '1' as a '0'. It follows that if p_j is the probability of allele 1 at marker j, then the probability of actually observing allele 1 will be given by $p_j^*=p_j(1-\epsilon_1)+(1-p_j)\epsilon_0$, which does not equal p_j in general, hence creating a bias. Under linkage equilibrium, the true haplotype probabilities that we wish to estimate are $P(y)=\prod p_j^{y_j}(1-p_j)^{1-y_j}$ for all L-tuples $y=(y_1,y_2,\ldots,y_L)$ of 0's and 1's, but the observed data actually follow the distribution $P^*(y)=\prod p_j^{*y_j}(1-p_j^*)^{(1-y_j)}$. Neglecting genotyping error, we are effectively estimating $P^*(y)$ rather than $P(y)$, leading to a bias of $b(y)=P^*(y)-P(y)$, and the asymptotic variance of the estimator will have the same form as before, but evaluated at $P^*(y)$ rather than $P(y)$. To take bias into account, we should define the efficiency of pooling relative to no pooling by the ratio of the sums of mean squared errors rather than the sums of variances. Thus (4) should be replaced by

$$\frac{\sum_{y\in\Psi}\{b^2(y)+V_1^*(y)\}}{\sum_{y\in\Psi}\{b^2(y)+V_K^*(y)\}}=\frac{\sum_{y\in\Psi}b^2(y)+\sum_{y\in\Psi}V_1^*(y)}{\sum_{y\in\Psi}b^2(y)+\sum_{y\in\Psi}V_K^*(y)},$$

where we have used $V_K^*(y)$ to denote $V_K(y)$ evaluated at p^*. To illustrate, we assume $\epsilon_0=\epsilon_1=0.05$ and obtain Table 4 instead of Table 2 for the efficiencies of pooling relative to no pooling when $n_P=50$. Comparing Table 4 with Table 2, we can see that genotyping error changes the relative efficiencies only slightly and the change is toward 1 because the squared bias term $\sum_{y\in\Psi}b^2(y)$ is common to both the numerator and denominator of the expression above. We will assume no genotyping error in the remainder of this article.

3.4 Efficiency under LD

While the preceding results are derived under linkage equilibrium, we can expect the same to hold true near linkage equilibrium. To gain some insights into what happens when there are moderate to high LD, we will carry out some efficiency calculation. To make it amenable to theoretical calculation without the need to resort to simulation studies, we consider the case where the loci can be grouped into blocks of size 2 each, with independence between blocks, and with intra-block LD coefficient $D'=0,0.25,0.5,0.75,0.99$. The minor allele frequencies are the same within blocks and equally spaced from 0.1 to 0.3 between blocks. We denote these block size 2 models by 2–2, 2–2–2 and so on. Thus, under model 2–2, the probability of the occurrence of haplotype (1,1,1,1) is $\{0.1^2+D'(0.1)(0.9)\}\{0.3^2+D'(0.3)(0.7)\}$. In general, $D'=0$ corresponds to the case of $L=2B$ independent loci, where B is the number of blocks, and the effective number of independent loci reduces from $2B$ to B when $D'=0.99$. By fixing the block size at 2, we are able to express the expected Fisher's information matrix in terms of certain quantities involving the non-central hypergeometric distribution (Xu *et al.*, 2008) which can be calculated accurately (Liao and Rosen, 2001). In Appendix A (Section A3), the derivation of the asymptotic variance–covariance matrix of the MLE of haplotype frequencies for model 2–2 is given. The derivations for models 2–2–2 and 2–2–2–2 can be similarly obtained.

The efficiencies of pooling relative to no pooling for estimating haplotype frequencies under these block size 2 models are reported in Table 5. Since many haplotypes have negligible probabilities of occurring under high LD, it may not be reasonable to include the variances of their estimates in the sums appearing in the numerator and denominator of (4). As an alternative measure, we consider thresholding the sums by including $V_1(y)$ and $V_K(y)$ in the sums only when $P(Y=y)$ exceeds a threshold. We consider a haplotype probability to be lower than the threshold if $P(Y=y)\le c/2^L$ for two choices of c, 0 and 0.1. The use of $c=0$ corresponds to no thresholding, whereas the choice $c=0.1$ would exclude a haplotype from the summations in (4) if its probability is less than or equal to one-tenth of the uniform probability $1/2^L$ for the case of L loci. It can be seen from Table 5 that the effect of thresholding on the resulting relative efficiencies are minimal and so we will concentrate on the non-threshold version in the following discussion.

When $D'=0$, all the loci are independent, which is why the first rows for models 2–2, 2–2–2 and 2–2–2–2 are close to the '$L=4$', '$L=6$' and '$L=8$' rows of Table 2. They are close but not identical because the marginal minor allele frequencies are different. To be specific, model 2–2 in Table 5 corresponds to the case $p_1=p_2=0.1$ and $p_3=p_4=0.3$, whereas the four loci case of Table 2 is for $p_1=0.1, p_2=0.167, p_3=0.233$ and $p_4=0.3$. We observe also that the last rows under models 2–2, 2–2–2 and 2–2–2–2 in Table 5 are very close to the '$L=2$', '$L=3$' and '$L=4$' rows of Table 2. This verifies our intuition that the B-block model reduces to the case of B independent loci when D' is large. Another observation from Table 5 is that the efficiency of pooling increases with D', which underlies why high LD is sometimes used to justify pooling. This phenomenon can again

Table 5. ARE of pooling relative to no pooling in estimating haplotype frequencies for various combinations of D' (LD coefficient) and K (pool size) for models 2–2, 2–2–2, and 2–2–2–2 with thresholding constant c

D'	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
Model 2–2				$c=0$				
0	1	0.869	0.674	0.522	0.412	0.332	0.272	0.227
0.25	1	0.951	0.771	0.611	0.492	0.390	0.325	0.277
0.5	1	1.168	1.077	0.947	0.822	0.706	0.608	0.524
0.75	1	1.435	1.618	1.671	1.655	1.605	1.532	1.453
0.99	1	1.594	1.988	2.267	2.477	2.639	2.768	2.874
				$c=0.1$				
0	1	0.899	0.717	0.563	0.450	0.366	0.306	0.260
0.25	1	0.968	0.802	0.655	0.518	0.433	0.357	0.305
0.5	1	1.221	1.181	1.072	0.953	0.835	0.742	0.651
0.75	1	1.488	1.725	1.829	1.846	1.824	1.778	1.713
0.99	1	1.597	1.993	2.275	2.486	2.650	2.781	2.889
Model 2-2-2				$c=0$				
0	1	0.412	0.184	0.092	0.051	0.030	0.019	0.012
0.25	1	0.478	0.228	0.119	0.067	0.041	0.026	0.017
0.5	1	0.679	0.403	0.244	0.154	0.100	0.068	0.048
0.75	1	0.953	0.781	0.619	0.488	0.385	0.307	0.248
0.99	1	1.157	1.125	1.049	0.967	0.891	0.822	0.762
				$c=0.1$				
0	1	0.443	0.202	0.107	0.060	0.038	0.024	0.016
0.25	1	0.516	0.261	0.143	0.084	0.053	0.034	0.023
0.5	1	0.713	0.447	0.281	0.183	0.124	0.087	0.061
0.75	1	0.985	0.831	0.676	0.544	0.439	0.357	0.292
0.99	1	1.161	1.131	1.056	0.974	0.898	0.830	0.769
Model 2–2–2–2				$c=0$				
0	1	0.192	0.049	0.016	0.006	0.003	0.001	0.001
0.25	1	0.236	0.067	0.022	0.009	0.004	0.002	0.001
0.5	1	0.383	0.145	0.061	0.028	0.014	0.008	0.004
0.75	1	0.617	0.363	0.219	0.137	0.088	0.058	0.040
0.99	1	0.820	0.612	0.462	0.357	0.283	0.229	0.194
				$c=0.1$				
0	1	0.205	0.055	0.019	0.008	0.004	0.002	0.001
0.25	1	0.253	0.076	0.027	0.011	0.005	0.003	0.001
0.5	1	0.409	0.165	0.074	0.035	0.018	0.010	0.006
0.75	1	0.647	0.397	0.249	0.161	0.108	0.074	0.051
0.99	1	0.824	0.618	0.467	0.362	0.287	0.233	0.195

be explained by the fact that the effective number of independent loci decreases as LD increases. The efficiency of pooling is always less than 1 in Table 5 for model 2–2–2–2 because the effective number of independent loci is at least 4. Pooling can be more efficient than no pooling for model 2–2–2, but only when $D'=0.99$, so that there are effectively only three independent loci, and when the pool size K is less than 5. The results for the case of intermediate LD are expectedly in between those for linkage equilibrium and high LD. The '$D'=0.75$' row under model 2–2 is peculiar in that the efficiency is not monotonic, but first increases with pool size and then decreases. An explanation of this unusual behavior is that according to Table 2, the efficiency decreases with pool size when the number of independent loci is 3, but increases with pool size when there are only two loci (which in turn can be explained by Table 1). When $D'=0.75$, the effective number of independent loci for model 2–2 is somewhere between 2 and 3 and hence the non-monotonic behavior.

4 DISCUSSION

We have shown in this article that contrary to the findings of Yang *et al.* (2003), pooling loses efficiency relative to no pooling in the estimation of haplotype frequencies when the number of independent loci is more than 3 and the alleles are not rare. Rare alleles cause sparsity which favors pooling. When the minor allele frequency is 0.05, pooling is more efficient than no pooling until the number of independent loci reaches 6 or 7. Rarer alleles will favor pooling even more. The effect of high LD is also to cause sparsity which allows pooling to maintain an advantage over no pooling for a larger number of loci than is possible under linkage equilibrium. To apply the guidelines derived under the assumption of linkage equilibrium, we find it useful to think in terms of the effective number of independent loci. For example, for model 2–2–2–2, the effective number of independent loci is 8 when $D'=0$ and 4 when $D'=0.99$. By interpolation, it seems reasonable to treat the cases $D'=0.25, 0.5, 0.75$ like there are 7, 6 and 5 independent loci, respectively, and this is substantiated by comparing Table 2 with Table 5.

To summarize, the main finding of this article is that for non-rare alleles and for the same genotyping cost, pooling loses efficiency relative to no pooling in the estimation of haplotype frequencies when the number of independent loci is more than 3. The critical number will increase for rare alleles, for example, to 6 when the minor allele frequency is 0.05. Our findings do not mean that pooling should not be applied in practice. There are circumstances under which pooling is more efficient than no pooling, namely, in situations when only a small number of haplotypes can occur with appreciable frequency (Barratt *et al.*, 2002) which can be caused by high LD, or rare alleles, or both. Thus, pooling is potentially useful in studies which employ dense sets of markers such as in the genome-wide context.

Table 4 suggests that genotyping error does not have a prominent effect on our findings. Ideally, one could incorporate genotyping error into the modeling and estimation procedure but this is beyond the scope of the present article. To compare haplotype frequencies in case–control studies, we must first estimate the frequencies separately for the cases and the controls and so our findings are still relevant.

Note that when we say no pooling in this article, we actually mean that we are not pooling individuals (i.e. pool size is 1). As commented by Xu *et al.* (2008), even when the pool size is 1, the data collected for each individual is typically genotype data which is a pool of two chromosomes or haplotypes. There will be information loss due to phase ambiguity. The ARE of haplotype estimation from individual genotype data without phase information relative to genotype data with phase information (which is much more expensive to get; easily 10 times more according to our colleagues who collect such data) can also be calculated using the results of this article for the asymptotic variance of the unphased data MLE and the multinomial variance formula for the phased data MLE. Assuming again that the minor allele frequencies are equally spaced between 0.1 and 0.3, the asymptotic efficiencies of the estimator based on unphased data relative to that based on phased data are $0.874, 0.717, \cdots, 0.101, 0.078$ for the case of 2–12

independent loci. Unlike in Tables 2 and 3, we have not taken cost into consideration so far. Suppose it is c times more expensive to obtain phased data, then the costs of collecting phased data for n_I individuals and unphased data for cn_I individuals are the same and the aforementioned asymptotic efficiencies should all be multiplied by c to give a fairer comparison based on the same cost. It follows that if $c=10$, there is no gain in estimation efficiency (for the same cost) to obtain phase information if the number of independent loci L is less than 12. The effect of LD is to reduce the effective number of independent loci and this will favor no phasing even more. This provides theoretical justification to the prevailing practice of not ascertaining phase information using molecular haplotyping method.

ACKNOWLEDGEMENTS

The authors are grateful to the three reviewers for their valuable suggestions which lead to improvements of the article, and Dr Yik Ying Teo for providing information on the cost of molecular phasing.

Conflict of Interest: none declared.

REFERENCES

Bansal,A. *et al*. (2002) Association testing by DNA pooling: an effective initial screen. *Proc. Natl Acad. Sci. USA*, **99**, 16871–16874.

Barcellos,L.F. *et al*. (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.*, **61**, 734–747.

Barratt,B.J. *et al*. (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.*, **66**, 393–405.

Fitzmaurice,G.M. *et al*. (1993) Regression models for discrete longitudinal responses (with discussion). *Stat. Sci.*, **8**, 284–309.

Homer,N. *et al*. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, **4**, e1000167.

Ito,T. *et al*. (2003) 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.*, **72**, 384–398.

Jacobs,K.B. *et al*. (2009) A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.*, **41**, 1253–1257.

Kirk,K.M. and Cardon,L.R. (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur. J. Hum. Genet.*, **10**, 616–622.

Kirkpatrick,B. *et al*. (2007) HAPLOPOOL: improving haplotype frequency estimation through DNA pools and phylogenetic modeling. *Bioinformatics*, **23**, 3048–3055.

Kuk,A.Y.C. *et al*. (2009) Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium. *Bioinformatics*, **25**, 379–386.

Liang,K.Y. *et al*. (1992) Multivariate regression analysis for categorical data (with Discussion). *J. R. Stat. Soc. B*, **54**, 3–40.

Liao,J.G. and Rosen,O. (2001) Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution. *Am. Stat.*, **55**, 366–369.

McLachlan,G.J. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.

Niu,T. (2004) Algorithms for inferring haplotypes. *Genet. Epidemiol.*, **27**, 334–347.

Norton,N. *et al*. (2004) DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.*, **36**, 146–152.

Pearson,J.V. *et al*. (2007) Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am. J. Hum. Genet.*, **80**, 126–139.

Quade,S. *et al*. (2005) Estimating haplotype frequencies in pooled DNA samples when there is genotyping error. *BMC Genet.*, **6**, 1471–2156.

Sham,P. *et al*. (2002) DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.*, **3**, 862–871.

Wang,S. *et al*. (2003) On the use of DNA pooling to estimate haplotype frequencies. *Genet. Epidemiol.*, **24**, 74–82.

Xu,J. *et al*. (2008) Testing linkage disequilibrium from pooled DNA: a contingency table perspective. *Stat. Med.*, **27**, 5801–5815.

Yang,Y. *et al*. (2003) Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl Acad. Sci. USA*, **100**, 7225–7230.

Zhang,H. *et al*. (2008) PoooL: an efficient method for estimating haplotype frequencies from large DNA pools. *Bioinformatics*, **24**, 1942–1948.

Zou,G. and Zhao,H. (2003) Haplotype frequency estimation in the presence of genotyping errors. *Hum. Hered.*, **56**, 131–138.

Zuo,Y. *et al*. (2006) Two-stage designs in case-control association analysis. *Genetics*, **173**, 1747–1760.

APPENDIX A

A1 Derivation of the asymptotic variance of the pooled data MLE of $\psi_{12\cdots L}$.

Since $\psi_{12\cdots L}=\alpha^{(L)}_{12\cdots L}$, we can work with the canonical parameters and the problem reduces to finding the asymptotic variance of the MLE of $\alpha^{(L)}_{12\cdots L}$. For a pool of K individuals, there are $n=2K$ DNA strands. Let $\mathbf{Y}_1=(Y_{11},\ldots,Y_{1L}),\ldots,\mathbf{Y}_n=(Y_{n1},\ldots,Y_{nL})$ be the haplotypes of the n strands. Assuming Hardy–Weinberg equilibrium, $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ are independent and identically distributed according to the distribution given by (1). With DNA pooling, we observe only $\mathbf{T}=\mathbf{Y}_1+\cdots+\mathbf{Y}_n=(T_1,\ldots,T_L)$, where $T_j=Y_{1j}+\cdots+Y_{nj}$ is the total allele frequency at locus j. The likelihood function based on the observed data $\mathbf{T}=\mathbf{Y}_1+\cdots+\mathbf{Y}_n$ can be obtained from the probability function (1) using the multivariate convolution formula and we can differentiate the resulting log-likelihood function to obtain the score function. This is very tedious and an easier way to obtain the score functions based on the observed data $\mathbf{T}=\mathbf{Y}_1+\cdots+\mathbf{Y}_n$ is to take conditional expectation of the score functions based on the unobserved data $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$, which is a well-known result in the EM literature (see, e.g. McLachlan and Krishnan, 1997, p. 100). It is straightforward to write down the score functions based on $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ because the distribution in (1) belongs to an exponential family. Taking conditional expectations of these exponential family score functions will give us the score functions based on the observed data $\mathbf{T}=\mathbf{Y}_1+\cdots+\mathbf{Y}_n$. Under linkage equilibrium, the score functions can be further simplified to

$$
\begin{aligned}
S_1&=T_1-np_1,\\
&\vdots\\
S_L&=T_L-np_L,\\
&\vdots\\
S_{12}&=n\frac{T_1}{n}\frac{T_2}{n}-np_1p_2,\\
&\vdots\\
S_{L-1,L}&=n\frac{T_{L-1}}{n}\frac{T_L}{n}-np_{L-1}p_L,\\
S_{123}&=n\frac{T_1}{n}\frac{T_2}{n}\frac{T_3}{n}-np_1p_2p_3,\\
&\vdots\\
S_{12\cdots L}&=n\frac{T_1}{n}\frac{T_2}{n}\cdots\frac{T_L}{n}-np_1p_2\cdots p_L.
\end{aligned}
$$

The last score $S_{12\cdots L}$ above corresponds to $\alpha^{(L)}_{12\cdots L}$ which is our focus. A simple way to calculate the asymptotic variance of the

MLE $\hat{\alpha}^{(L)}_{12\cdots L}$ of $\alpha^{(L)}_{12\cdots L}$, as the number of pools n_P increases, is to find the efficient score which is the projection of the score $S_{12\cdots L}$ to the space orthogonal to all the other scores. When this is done, the asymptotic variance of $\hat{\alpha}^{(L)}_{12\cdots L}$ is n_P^{-1} times the reciprocal of the variance of the efficient score. It can be easily verified that the efficient score is

$$S = n\frac{(T_1 - np_1)}{n}\frac{(T_2 - np_2)}{n}\cdots\frac{(T_L - np_L)}{n}$$

since it can be written as $S_{12\cdots L}$ plus a linear combination of the other scores, and it is orthogonal/uncorrelated to all the other scores as a result of centering and independence. Using the facts that $E(S)=0$ and the T_j's are independent under linkage equilibrium, the variance of the efficient score S is

$$\mathrm{var}(S) = E(S^2) = n^2\prod_{j=1}^{L}\frac{np_j(1-p_j)}{n^2} = n^{2-L}\prod_{j=1}^{L}p_j(1-p_j)$$

from which (2) follows.

A2 Derivation of the asymptotic variance–covariance matrix of the pooled data MLE of the canonical parameters α's.

An orthogonal basis for the 2^L-1 score functions $S_1,\ldots,S_L,S_{12},\ldots,S_{L-1,L},S_{123},\ldots,S_{L-2,L-1,L},\ldots,S_{12\cdots L}$ of the canonical parameters is given by

$$Z_1 = S_1 = T_1 - np_1,$$
$$\vdots$$
$$Z_L = S_L = T_L - np_L,$$
$$\vdots$$
$$Z_{12} = n\frac{(T_1 - np_1)}{n}\frac{(T_2 - np_2)}{n},$$
$$\vdots$$
$$Z_{L-1,L} = n\frac{(T_{L-1} - np_{L-1})}{n}\frac{(T_L - np_L)}{n},$$
$$Z_{123} = n\frac{(T_1 - np_1)}{n}\frac{(T_2 - np_2)}{n}\frac{(T_3 - np_3)}{n},$$
$$\vdots$$
$$Z_{12\cdots L} = n\frac{(T_1 - np_1)}{n}\frac{(T_2 - np_2)}{n}\cdots\frac{(T_L - np_L)}{n}.$$

One can readily express the scores $S_1,\ldots,S_L,S_{12},\ldots,S_{L-1,L},S_{123},\ldots,S_{L-2,L-1,L},\ldots,S_{12\cdots L}$ as linear combinations of the Z's by taking inner products. As a result, we can obtain the Fisher's information matrix, which is just the variance–covariance matrix of the scores, from the variance–covariance matrix of the Z's which is diagonal due to orthogonality. The asymptotic variance–covariance matrix of the MLE of the canonical parameters is n_P^{-1} times the inverse of the Fisher's information matrix.

A3 Derivation of the asymptotic variance–covariance matrix of the MLE of haplotype frequencies in model 2–2

For loci 1 and 2 in block 1, define $T_{12} = \sum_{i=1}^{n} Y_{i1}Y_{i2}$. We can again obtain the score vector by taking conditional expectation of the 'complete data' score vector. Let

$$\tilde{T}_{12} = E[T_{12}|T_1,T_2],$$
$$n_1^* = (T_1, T_2, \tilde{T}_{12})^T,$$
$$p_1^* = (p_1, p_2, p_{12})^T,$$
$$s_1 = n_1^* - np_1^*,$$

and

$$I_1 = \begin{Bmatrix} np_1(1-p_1) & n(p_{12}-p_1p_2) & n(p_{12}-p_1p_{12}) \\ n(p_{12}-p_1p_2) & np_2(1-p_2) & n(p_{12}-p_2p_{12}) \\ n(p_{12}-p_1p_{12}) & n(p_{12}-p_2p_{12}) & \mathrm{var}(\tilde{T}_{12}) \end{Bmatrix}.$$

Define $\tilde{T}_{34}, n_2^*, p_2^*, I_2$ similarly for loci 3 and 4 in block 2. Exploiting intra-block independence and define

$$s_{12} = \frac{n_2^* \otimes n_1^*}{n} - np_2^* \otimes p_1^*,$$

$$I = \begin{Bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & \frac{I_2 \otimes I_1}{n^2} \end{Bmatrix}, A = \begin{Bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ p_2^* \otimes I_1 & I_2 \otimes p_1^* & \frac{I_2 \otimes I_1}{n^2} \end{Bmatrix},$$

then the score vector for $\alpha = (\alpha_1, \alpha_2, \alpha_{12}, \alpha_3, \alpha_4, \alpha_{34}, \alpha_{13}, \alpha_{14}, \alpha_{134}, \alpha_{23}, \alpha_{24}, \alpha_{234}, \alpha_{123}, \alpha_{124}, \alpha_{1234})^T$ is $s = (s_1^T, s_2^T, s_{12}^T)$. By utilizing blockwise orthogonalization techniques, its covariance matrix is $AI^{-1}A^T$. Note that conditional on T_1 and T_2, T_{12} follows a hypergeometric distribution and

$$\mathrm{Var}(\tilde{T}_{12}) = n * p_{12} * (1-p_{12}) - E[\mathrm{Var}(T_{12}|T_1,T_2)].$$

As suggested by Liao and Rosen (2001), $E[\mathrm{Var}(T_{12}|T_1,T_2)]$ can be accurately evaluated by generating a large random sample of T_1, T_2 and then averaging $\mathrm{Var}(T_{12}|T_1,T_2)$ across the sample. Denote the haplotype frequencies by $p = \{p_{y_1y_2y_3y_4}, y_i = 0,1, 1 \le i \le 4.\}$. We have $\frac{\partial p_{y_1y_2y_3y_4}}{\partial \alpha_{i_1\ldots i_k}} = p_{y_1y_2y_3y_4}(y_{i_1}\ldots y_{i_k} - p_{i_1\ldots i_k})$, where $p_{i_1\ldots i_k}$ is the probability of allele '1' at loci $i_1,\ldots,i_k$. Denote $\frac{\partial p}{\partial \alpha}_{16\times 15}$ by B. By the delta method, the asymptotic variance–covariance matrix for the MLE of p is $B(AI^{-1}A^T)^{-1}B^T$.