Stony Brook University

Hardy-Weinberg Deviation and EM-based Haplotype Frequency Estimation

A Dissertation Presented

by

Hyeong Jun Ahn

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philsophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2011

Stony Brook University

The Graduate School

Hyeong Jun Ahn

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

John J. Chen, Associate Professor
Departments of Preventive Medicine & Applied Mathematics and Statistics

Nancy R. Mendell, Professor
Department of Applied Mathematics and Statistics

Wei Zhu, Professor
Department of Applied Mathematics and Statistics

Barbara Nemesure, Associate Professor
Department of Preventive Medicine

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Hardy-Weinberg Deviation and EM-based Haplotype Frequency Estimation

by

Hyeong Jun Ahn

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2011

Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variation in human genome. Haplotypes which combine multiple SNPs into super-alleles have been widely used in modern genetic analysis, especially in human disease association studies. The Expectation Maximization (EM) algorithm is commonly used in haplotype phasing and frequency estimation, and Hardy-Weinberg (HW) equilibrium is a key assumption built into the EM algorithm. The accuracy of EM-based haplotype frequency estimation when the HW equilibrium assumption is violated has been explored by several studies. The general consensus is that the sampling error plays a more dominant role in haplotypes estimation than the estimation error due to HW deviation; the accuracy of haplotype frequency estimation tends to improve with increasing homozygosity in the sample. However, these studies mainly concentrated on the impact of SNP level HW deviation. A theoretical foundation for the impact of HW deviation at the haplotype level on haplotype frequency estimation has not been established.

In this dissertation, we derived the theoretical relationship among three haplotype mean squared errors: between population and sample frequencies (MSE_{PS}), between true sample and sample estimated frequencies (MSE_{SE}), and between population and sample estimated frequencies (MSE_{PE}). The theoretical relationship between SNP level and haplotype level HW deviations was also established. Our simulations show that the violation of HW equilibrium at haplotype level could result in more severe haplotype estimation error than sampling error, and the accuracy of haplotype frequency estimation is not always improved with increasing homozygosity.

To incorporate the possible haplotype level HW deviations into the haplotype frequency estimation process, we propose a Hardy-Weinberg Deviation-Expectation/Conditional Maximization (HWD-ECM) method which allows us to estimate HW deviation parameters and haplotype frequencies simultaneously. For two SNPs cases, the HWD-ECM algorithm consists of three iteration steps: 1). an expectation step estimating genotype frequencies allowing HW deviation parameters; 2). a conditional maximization step for HW deviation parameter estimation utilizing constraints of SNP level or haplotype level HW deviation parameters; and 3). a conditional maximization step for haplotype frequencies. Simulation results show that the HWD-ECM method performs significantly better than the EM-based approach in haplotype estimation when HWE assumption is violated. Algorithm for extension of HWD-ECM to multiple SNPs is also discussed.

Table of Contents

List of Figures

List of Tables

Acknowledgment

First and foremost I want to thank my advisor, Professor John J. Chen, for his great encouragement and guidance in the past five years. He has been a role model for my entire graduate studying time and it has been an honor to be his Ph.D. student. I also thank my committee members, Professor Nancy R. Mendell, Professor Wei Zhu and Professor Barbara Nemesure. Their advice and help were indispensable for my dissertation and other research projects.

I thank AMS department for providing me with an excellent academic environment. I also appreciate the help and friendship from my officemate Guangxiang Zhang who shows a good example of a Ph.D. student.

I thank my wife, Jinah Kim, for her prayers and support during my academic pursuit. Without her love and patience, I wouldn't have been here. I also thank my daughter, Hannah Ahn who was born in 2010, for giving me energy to continue.

I thank Jesus for being a reason that I live for and hope that all my works glorify God.

Chapter 1

Introduction

Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variation in human genome. About 10 million SNPs exist in human populations for which the rarer SNP allele has a frequency of at least 1 percent (Sobrino et al. 2005). Alleles of SNPs that are close to each other tend to be inherited together. A set of associated SNP alleles in a region of a chromosome is called a haplotype. The haplotype block formed by these associated SNPs has very valuable information in detecting genes or region causing common diseases. Haplotype plays a key role in genetic association studies since haplotype block structure in human genome is related to hot spots and cold spots for recombination (Daly et al. 2001; Fallin et al. 2001; Arnheim et al. 2003; Schaid 2004). A haplotype map, or HapMap, intended to reveal such variation patterns, has been recently developed by the International HapMap Consortium (The International HapMap 2005). Once such variants have been discovered, we can learn much more about the origins of illnesses and about ways to prevent, diagnose, and treat those illnesses.

However, haplotype usually cannot be obtained directly from unphased genotype data. Molecular experimental techniques were developed, including single-molecule dilution (Stephens et al. 1990), long-range allele-specific PCR (MichalatosBeloin et al. 1996), diploid-to-haploid conversion (Douglas et al. 2001), carbon nanotube probing (Woolley et al. 2000), but such methods are not widely used because they are too expensive and low-throughput at this time for population research. Haplotypes can also be resolved via family data, which is also

expensive to collect (Wijsman 1987). Therefore, haplotype determination through statistical methods is most commonly used. Clark's algorithm is the first statistical method for haplotype frequency estimation from genotypes of unrelated individuals (Clark 1990), but more sophisticated methods such as maximum likelihood methods or Bayesian approaches have been developed.

Bayesian algorithm, incorporating prior information into the statistical model, has been applied to haplotype frequency estimation. Stephens et al. (2001) proposed a coalescence-based Markov-chain Monte Carlo (MCMC) approach. Instead of using a prior based on the coalescence theory, a Dirichlet prior was also used in the Gibbs sampling (Niu et al. 2002). Stephens and Donnelly (2003) modified the coalescence-based MCMC approach by incorporating a variant of the partition-ligation idea and by allowing for recombination and decay of linkage disequilibrium (LD) with distance. Since Bayesian algorithms depend on the prior information, whether the algorithm performs favorably compared to other algorithms when the prior model does not hold remains to be seen (Niu 2004).

The Expectation Maximization (EM) algorithm (Dempster et al. 1977), a maximum likelihood based method, is commonly used in haplotype frequency estimation. The earlier works were first introduced to estimate haplotype frequencies from unrelated individuals (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). Partition-Ligation approach was developed as a new strategy to infer haplotypes with large number of SNPs using the EM algorithm (Qin et al. 2002). Li et al. applied the estimation equation technique and further improved the statistical and computational efficiency in the estimation of haplotype frequencies (Li et al. 2003). EM algorithm was also used to estimate haplotype frequencies based on pedigree data (Zhang et al. 2006; Zhu et al. 2007). EM-based method was also developed to

reduce the impact on haplotype estimation of genotyping errors (Zhu et al. 2009). EM based methods have the advantage of being prior model-free, but all EM approaches assume Hardy-Weinberg equilibrium in their algorithms.

Hardy-Weinberg equilibrium (HWE) was independently introduced first by Hardy and Weinberg in early 1900s (Hardy 1908, Weinberg 1908). Within the EM algorithm HWE assumption allows the replacement of genotype frequencies by the product of haplotype frequencies. There were several attempts to consider HW deviation in haplotype analysis. Single et al. (2002) tried to improve the accuracy of haplotype frequency estimation by removing some loci which showed significant SNP level departure from HWE. However, their method didn't consider the impact of haplotype level HW deviation. Epstein and Satten (2003) used an Expectation/Conditional Maximization (ECM) (Meng and Rubin 1993) approach for inference of haplotype effects in a genetic association study setting (Epstein and Satten 2003). They attempted to add a common fixation index (F) to allow some deviation from HWE (Satten and Epstein 2004). The method was applicable to case control studies only and allowed only a single fixation index. Kuk et al. (2009) developed a method to estimate haplotype frequencies from pooled DNA with or without HWE assumption (Kuk et al. 2009). However, pooling genotype has disadvantages such as loss of individual genotype information and relatively high measurement error (Zhang et al. 2008).

Fallin and Schork (2000) investigated the accuracy of haplotype frequency estimation through simulation studies. They compared the mean square error between population haplotype frequencies and estimated haplotype frequencies, and the mean square error between sample haplotype frequencies and estimated haplotype frequencies. They simulated locus-specific allelic departures from Hardy-Weinberg equilibrium, i.e., SNP level HW deviations. They concluded

that the majority of errors between population haplotype frequencies and estimated haplotype frequencies were caused by the sampling error, not the estimation error. In addition, they concluded that HW deviation toward excessive heterozygosity increased estimation errors, and HW deviation toward excessive homozygosity improved the estimation accuracy.

In this dissertation, we establish the theoretical relationship between SNP level and haplotype level HW deviations and among the three haplotype mean square errors (MSEs). We investigate how haplotype level HW deviations impact various MSE measures of haplotype frequency.

To reduce haplotype frequency estimation error, HW deviation parameters need to be incorporated in the estimation process. We propose a Hardy-Weinberg Deviation-Expectation/Conditional Maximization (HWD-ECM) method which allows us to estimate HW deviation parameters and haplotype frequencies simultaneously. HWD-ECM algorithm is an extension of ECM algorithm (Meng and Rubin 1993) which has two conditional maximization steps rather than a single complicated joint maximization step.

The dissertation is organized as follows. In Chapter 2, we first review Hardy-Weinberg equilibrium, the standard EM approach and several measures of assessing the accuracy of haplotype frequency estimation. The relationship between SNP and haplotype level HW deviations is established and theoretical bounds of haplotype level HW deviation are derived in Chapter 3. Simulation results of the impact of HW deviations on haplotype frequency estimation for two SNPs scenario are also summarized. In Chapter 4, the HWD-ECM approach for haplotypes and HW deviations' simultaneous estimation is proposed and its advantage over traditional EM approach is established through simulation studies. The extension of HWD-ECM

approach for multiple SNPs is also described. Discussions about HW deviations, haplotype frequency estimation and future work are provided in Chapter 5.

Chapter 2

Hardy-Weinberg equilibrium and standard EM method for haplotype estimation

2.1. Hardy-Weinberg equilibrium (HWE)

For a single locus with two alleles A or a, we denote allele frequency of allele A by p and of allele a by q. If the population is in equilibrium under conditions of no mutation, no gene flow, no genetic drift, random mating and no natural selection, then we will have genotype frequency $P(AA) = p^2$ for the AA homozygotes, $P(aa) = q^2$ for the aa homozygotes, and $P(Aa) = 2pq$ for the heterozygotes in the population. Procedures for testing HWE have been extensively investigated (Elston and Forthofer 1977; Emigh 1980; Hernandez and Weir 1989; Guo and Thompson 1992; Gomes et al. 1999; Cox and Kraft 2006). The modern concept of Hardy-Weinberg (HW) deviation was introduced by Hernandez and Weir in 1989. For a locus with two alleles, A and B, HW deviation parameters are defined as

$$D_{AA} = P(AA) - p^2, D_{aa} = P(aa) - q^2, D_{Aa} = P(Aa) - 2pq.$$

HW equilibrium can be similarly defined for multiple alleles and applied to haplotype blocks. Statistical tests have been developed to directly evaluate the hypotheses in term of HW deviation parameters (Hernandez and Weir 1989; Chen and Thomson 1999; Chen et al. 2005)

2.2. Expectation Maximization (EM) approach

The EM algorithm is an iterative method for finding maximum likelihood (Dempster et al. 1977) and it is the leading numerical methods used to obtain the maximum likelihood estimation of haplotypes. The EM algorithm consists of an expectation step and a maximization step, computing the sets of haplotype frequencies, $p_1, p_2, \cdots, p_h$, iteratively starting with an initial set of values, $p_1^{(0)}, p_2^{(0)}, \cdots, p_h^{(0)}$ (Excoffier and Slatkin 1995). The observed unphased genotype frequencies ($n_1, n_2, \cdots, n_m$) follow a multinomial distribution with unphased genotype probabilities, $P_1, P_2, \cdots, P_m$:

$$P_j = \sum_{i=1}^{c_j} P(h_k h_l)_i \ ,$$

where c_j is the number of phased genotypes leading to the j[th] unphased genotype and $P(h_k h_l)_i$ is the probability of the i[th] phased genotype made up of haplotypes k and l. In the current dissertation, we will also use the term phenotype interchangeably with unphased genotype.

Expectation step

At the expectation step, initial haplotype frequency values are used to estimate the genotype frequencies of the 1[st] iteration. The g^{th} iteration of genotype frequency, $\hat{P}(h_k h_l)^{(g)}$, can be estimated as follows:

$$\hat{P}(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{P(genotype\ h_k h_l\ in\ phenotype\ j)}{P(phenotype\ j)} = \frac{n_j}{n} \frac{P_j(h_k h_l)^{(g)}}{P_j{}^{(g)}},$$

Assuming HWE, $P_j\,(h_k h_l)^{(g)} = \begin{cases} (p_k^{(g)})^2, & if\ k = l, \\ 2p_k^{(g)} p_l^{(g)}, & if\ k = l, \end{cases}$ where $p_k^{(g)}$ is the g^{th} iteration of frequency of haplotype *k* frequency and $p_l^{(g)}$ is the g^{th} iteration of frequency of haplotype *l*.

Homozygous or single heterozygous genotypes can be phased directly without error and only multiple heterozygous genotypes (double heterozygous genotype for two SNPs scenarios) need to be estimated through the above expectation step.

Maximization step

The expectation step's estimated genotype frequencies from the current iteration are used to update the next iteration's haplotype frequencies in the maximization step. Given all phased genotype frequencies, the complete log likelihood is as follows:

$$\log(L_c) = \sum_{(k,l)\in h} n_{kl} \log(P(h_k h_l))\,.$$

The maximum likelihood estimation of haplotype frequencies can then be computed easily by taking partial derivative of the complete log likelihood with respect to each haplotype frequency and the t th haplotype frequency of (g+1) th iteration is calculated as follows:

$$p_t{}^{(g+1)} = P(h_t h_t)^{(g)} + \frac{1}{2} \sum_{i=1, i\neq t}^{h} P(h_t h_i)^{(g)}.$$

The above equation is equivalent to the gene counting method. The EM iteration will stop under predetermined convergence criteria based on absolute change in log likelihood $(e.g.,< 10^{-6}$ for R package haplo.em).

2.3. Accuracy measures for haplotype frequency estimation

2.3.1. Mean squared errors for haplotype

The mean squared error (MSE) for haplotype is our primary measure of accuracy. Three MSEs can be defined among the haplotype frequencies: population haplotype frequencies, true sample haplotype frequencies and sample estimated haplotype frequencies, i.e. MSE_{PE} (mean squared error between population and estimated haplotype frequencies), MSE_{PS} (mean squared error between population and sampled haplotype frequencies) and MSE_{SE} (mean squared error between sampled and estimated haplotype frequencies). In a phased genotype sample, the sampled haplotype frequency is conventionally calculated by counting the number of each haplotype. Since each individual has one haplotype pair, the total number of haplotype should be twice as many as the total number of individuals. Fallin and Schork (2000) treated [MSE_{PE}-MSE_{SE}] as the sampling error of haplotype frequencies and compare it with $MSE_{SE,}$ which was the estimation error. In the following Theorem, we derive the theoretical relationship among three MSEs.

Theorem 2.1. *Let $(P_t)_{POP}$ be the t th haplotype frequency of the population and $(P_t)_{SAM}$ be the sample haplotype frequency and $(P_t)_{EM}$ be the estimated haplotype frequency from EM algorithm, then the mean squared errors (MSEs) have the following relationship:*

$$MSE_{PE} = MSE_{PS} + MSE_{SE} + 2\frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM}) \qquad (2.1)$$

Proof.

$$MSE_{PE} = \frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{EM})^2 = \frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM} + (P_t)_{SAM} - (P_t)_{EM})^2$$

$$= \frac{1}{h}\sum_{t=1}^{h}\{((P_t)_{POP} - (P_t)_{SAM})^2 + 2((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM}) + ((P_t)_{SAM} - (P_t)_{EM})^2\}$$

$$= \frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})^2 + 2\frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM}) + \frac{1}{h}\sum_{t=1}^{h}((P_t)_{SAM} - (P_t)_{EM})^2$$

$$= MSE_{PS} + MSE_{SE} + 2\frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM}).$$

■

Besides MSE_{PS} and MSE_{SE}, there is an additional term $2\frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM})$ in(2.1). Therefore,$[MSE_{PE} - MSE_{SE}]$, which was used as sampling error by Fallin and Schork (2000) is actually $MSE_{PS} + 2\frac{1}{h}\sum_{t=1}^{h}((P_t)_{POP} - (P_t)_{SAM})((P_t)_{SAM} - (P_t)_{EM})$. In the following MSE_{PS} will be sampling error and MSE_{SE} used as estimation error, and MSE_{PE} as the total MSE.

For two SNPs scenario, since there is no estimation error for homozygous or single heterozygous genotype, MSE_{SE} is affected solely by the difference between sample and sample estimated double heterozygous genotypes.

Theorem 2.2. *For a two locus situation, let $P(h_t h_i)_{SAM}^{double}$ be the sample double heterozygous genotype frequency with corresponding haplotypes h_t and h_i, $P(h_t h_i)_{EM}^{double}$ be the EM estimated double heterozygous genotype frequency, $(P_i)_{SAM}$ and $(P_t)_{SAM}$ be the sample haplotype frequencies, $(P_i)_{EM}$ and $(P_t)_{EM}$ be the EM estimated haplotype frequencies, then the estimation MSE can be described as*

$$MSE_{SE} = \frac{1}{h}\sum_{t=1}^{h}(\frac{1}{2}\sum_{i=1,i\neq t}^{h}\{P(h_t h_i)_{SAM}^{double} - P(h_t h_i)_{EM}^{doub\ le}\})^2$$

Proof.

$(P_t)_{SAM} - (P_t)_{EM}$

$$= P(h_t h_t)_{SAM} + \frac{1}{2}\sum_{i=1,i\neq t}^{h} P(h_t h_i)_{SAM} - \{P(h_t h_t)_{EM} + \frac{1}{2}\sum_{i=1,i\neq t}^{h} P(h_t h_i)_{EM}\}$$

$$= \frac{1}{2}\sum_{i=1,i\neq t}^{h}\{P(h_t h_i)_{SAM}^{double} - P(h_t h_i)_{EM}^{double}\},$$

Therefore, for a two SNPs system, the estimation error for haplotype is affected by only double heterozygous genotypes as

$$MSE_{SE} = \frac{1}{h}\sum_{t=1}^{h}((P_t)_{SAM} - (P_t)_{EM})^2 = \frac{1}{h}\sum_{t=1}^{h}(\frac{1}{2}\sum_{i=1,i\neq t}^{h}\{P(h_t h_i)_{SAM}^{double} - P(h_t h_i)_{EM}^{double}\})^2$$

■

When the frequency for double heterozygous genotypes is estimated incorrectly, the haplotype estimation error would occur.

2.3.2. Mean squared errors for genotype

Similar to the above discussion on MSE_{PS} for haplotype, the sampling error for genotype can be defined as

$$MSE_{PS}(genotype) = \frac{2}{h(h+1)}\sum_{i=1}^{h}\sum_{j=1,i\leq j}^{h}(P(h_i h_j)_{POP} - P(h_i h_j)_{SAM})^2,$$

where total number of genotype can be counted as $\frac{h(h+1)}{2}$ with h being the total number of haplotypes.

Also, the estimation error for genotype can be defined as

$MSE_{SE}(genotype) = \frac{2}{h(h+1)}\sum_{i=1}^{h}\sum_{j=1,i\leq j}^{h}(P(h_t h_i)_{SAM}^{double} - P(h_t h_i)_{EM}^{double})^2$, where $P(h_t h_i)_{SAM}^{double}$ is the sample double heterozygous genotype frequency with corresponding

haplotypes h_t and h_i and $P(h_t h_i)_{EM}^{double}$ is the EM estimated double heterozygous genotype frequencies since there is no estimation error for homozygous and single heterozygous genotype.

2.4. Haplotype counting

According to Excoffier and Slakin (1995), calculation of haplotype frequencies from each maximization step is equivalent to the conventional gene-counting method (Ceppellini et al. 1955; Smith 1957). At this section, we show how the gene-counting method is derived by taking partial derivative of log likelihood with respect to haplotype. When genotypes are in HWE, the complete log likelihood of phased genotype frequencies is

$$\log(L_c) = \sum_{(k,l)\in h} n_{kl} \log(P(h_k h_l))$$

$$= n_{tt} \log(P_t^2) + \sum_{i=1,i\neq t} n_{ti} \log(2P_t P_i) + \sum_{i=1,i\neq t} \sum_{j=1,j\neq t} n_{ij} \log\left(2P_i P_j\right)$$

Since $\sum P_i = 1$, the Lagrange multiplier λ with its constraint should be considered to obtain the maximum likelihood of P_t. The Lagrange multiplier λ can be derived as -2n through simple algebraic calculations. After taking partial derivative of the above complete log likelihood with respect to P_t , we have

$$n_{tt} \frac{2P_t}{P_t^2} + \sum_{i=1,i\neq t} n_{ti} \frac{2P_i}{2P_t P_i} - 2n = n_{tt} \frac{2}{P_t} + \sum_{i=1,i\neq t} n_{ti} \frac{1}{P_t} - 2n = 0$$

When we solve the above equation for P_t , we can easily derive the maximization step result as $P_t = \frac{2n_{tt} + \sum_{i=1,i\neq t} n_{ti}}{2n} = P(h_t h_t) + \frac{1}{2}\sum_{i=1,i\neq t}^{h} P(h_t h_i)$.

Chapter 3

Impact of Hardy-Weinberg deviation on standard EM haplotype estimation

One key assumption in the expectation step of the EM estimation of the gametic phase of multiple heterozygous genotypes is HWE. When true genotype samples are deviated from HWE, the assumption in expectation step can potentially cause severe phasing error, resulting in large estimation error for haplotype and genotype. Fallin and Schork (2000) studied the impact of SNP level departure from HWE on haplotype estimation accuracy when alleles at the loci are not in HWE. What directly affects the haplotype accuracy is haplotype level HW deviations. How SNP level HW deviation relates to haplotype level HW deviation has not been established. In order to investigate the impact on estimation error, we need to first explore this relationship.

3.1. SNP level and haplotype level Hardy-Weinberg deviations

For two SNPs scenarios, we assume that the two alleles are A and a at the first locus and B and b at the second locus. We denote P_{AA} for the AA homozygotes, P_{aa} for the aa homozygotes, and P_{Aa} for the heterozygotes at first locus in the population, with allele frequencies of A and a can be denoted by p_A and p_a , respectively. Similarly, we can define P_{BB}, P_{bb} ,P_{Bb} , p_B and p_b for the second locus. One-locus genotype frequencies are defined by the following relations (Weir 1996):

$$P_{AA} = p_A^2 + D_{AA},\ P_{Aa} = 2p_A p_a + D_{Aa},\ P_{aa} = p_a^2 + D_{aa}\ , \qquad (3.1)$$

where D_{AA}, D_{Aa} and D_{aa} are the SNP level HW deviations corresponding to each genotypes at the first locus. Similarly we can define D_{BB}, D_{Bb} and D_{bb} for the second locus.

Given the four possible haplotypes (AB, Ab, aB, ab), genotype frequencies can then be defined by the haplotypes and haplotype level HW deviations in the following way:

$$P_{i|j} = \begin{cases} P_i^2 + D_{i|i}, & if\ i = j, \\ 2P_i P_j + D_{i|j}, & if\ i \neq j, \end{cases}$$

where $i\ and\ j\ \in (AB, Ab, aB, ab)$.

The relationship between SNP level and haplotype level HW deviations can be established by counting alleles or haplotypes from genotypes.

Theorem 3.1. *Let $D_{AA}, D_{Aa}\ and\ D_{aa}$ be SNP level HW deviations at the first locus, $D_{BB}, D_{Bb}\ and\ D_{bb}$ be SNP level HW deviations at the second locus, and $D_{i|j}$ be haplotype level HW deviations where $i\ and\ j\ \in (AB, Ab, aB, ab)$. Then,*

$$D_{AA} = D_{AB|AB} + D_{AB|Ab} + D_{Ab|Ab}$$

$$D_{Aa} = D_{AB|aB} + D_{AB|ab} + D_{Ab|aB} + D_{Ab|ab}$$

$$D_{aa} = D_{aB|aB} + D_{aB|ab} + D_{ab|ab}$$

$$D_{BB} = D_{AB|AB} + D_{AB|aB} + D_{aB|aB}$$

$$D_{Bb} = D_{AB|Ab} + D_{AB|ab} + D_{Ab|aB} + D_{aB|ab}$$

$$D_{bb} = D_{Ab|Ab} + D_{Ab|ab} + D_{ab|ab}$$

Proof.

The SNP level genotype frequency P_{AA} can be expressed by the sum of haplotype level genotypes sharing allele A:

$$\begin{aligned} P_{AA} &= P_{AB|AB} + P_{AB|Ab} + P_{Ab|Ab} \\ &= (P_{AB})^2 + D_{AB|AB} + 2P_{AB}P_{Ab} + D_{AB|Ab} + (P_{Ab})^2 + D_{Ab|Ab} \\ &= (P_{AB} + P_{Ab})^2 + D_{AB|AB} + D_{AB|Ab} + D_{Ab|Ab} \end{aligned}$$

Since the allele frequency p_A can be expressed as sum of P_{AB} and P_{Ab} ,

$$p_A^2 + D_{AB|AB} + D_{AB|Ab} + D_{Ab|Ab} = p_A^2 + D_{AA} \text{ by definition (3.1)}$$

Therefore, we can obtain $D_{AA} = D_{AB|AB} + D_{AB|Ab} + D_{Ab|Ab}$. Other formula can be similarly derived.

■

Based on Theorem 3.1., we have the following two corollaries.

Corollary 3.1. *Let D_{AA}, D_{Aa} and D_{aa} be SNP level HW deviations for first locus and D_{BB}, D_{Bb} and D_{bb} be SNP level HW deviations for second locus. Suppose $D_{i|j} = 0$, where i and $j \in (AB, Ab, aB, ab)$. Then*

$$D_{AA} = D_{Aa} = D_{aa} = D_{BB} = D_{Bb} = D_{bb} = 0.$$

Based on Theorem 3.1, we can conclude that if haplotype level HW deviations are zero (i.e. in HWE), then SNP level HW deviations is also zero (i.e. in HWE).

Corollary 3.2. *Let D_{AA}, D_{Aa} and D_{aa} be SNP level HW deviations for the first locus and D_{BB}, D_{Bb} and D_{bb} be SNP level HW deviations for the second locus. Suppose that $D_{i|j} \neq 0, where\ i\ and\ j \in (AB, Ab, aB, ab)$, $\sum_{l,n\in(B,b)} D_{kl|mn} = 0$ for k and $m \in (A, a)$ and $\sum_{k,m\in(A,a)} D_{kl|mn} = 0$ for l and $n \in (B, b)$. Then*

$$D_{AA} = D_{Aa} = D_{aa} = D_{BB} = D_{Bb} = D_{bb} = 0.$$

In other words, if SNP level HW deviations are zero, haplotype HW deviations are not necessarily zero. Furthermore, the SNP level HW deviations don't guarantee the extent of haplotype level HW deviation because SNP level HW deviations are sums of the haplotype level HW deviations corresponding to each SNP. Specifically, even though SNP level HW deviations are zero, haplotype level HW deviations can still be severe. This suggests that the impact of HW deviation on EM-based approach should be explored at haplotype level rather than at SNP level.

3.2. Constraints and bounds of SNP level and haplotype level Hardy-Weinberg deviations

SNP level HW deviations' constraints have been established (Weir 1996):

$$D_{AA} + \frac{1}{2}D_{Aa} = 0\ and\ D_{aa} + \frac{1}{2}D_{Aa} = 0.$$

Similarly we can establish the constraints of haplotype level HW deviations.

Theorem 3.2. *For two SNPs scenario, let* $D_{i|j}$ *be a haplotype level HW deviation where* $i \text{ and } j \in (AB, Ab, aB, ab)$. *Then,*

$$D_{i|i} + \frac{1}{2}\sum_{i \neq j} D_{i|j} = 0.$$

Proof.

This can be proof since haplotype frequencies can be obtained by genotypes related to each haplotype,

$$P_{AB} = P_{AB|AB} + \frac{1}{2}\left(P_{AB|Ab} + P_{AB|aB} + P_{AB|ab}\right)$$

$$= P_{AB}^2 + D_{AB|AB} + \frac{1}{2}\left(2P_{AB}P_{Ab} + D_{AB|Ab} + 2P_{AB}P_{aB} + D_{AB|aB} + 2P_{AB}P_{ab} + D_{AB|ab}\right)$$

$$= P_{AB}(P_{AB} + P_{Ab} + P_{aB} + P_{ab}) + D_{AB|AB} + \frac{1}{2}\left(D_{AB|Ab} + D_{AB|aB} + D_{AB|ab}\right) = P_{AB}$$

Then we have the constraint of HW deviation related to haplotype AB as

$$D_{AB|AB} + \frac{1}{2}\left(D_{AB|Ab} + D_{AB|aB} + D_{AB|ab}\right) = 0$$

For other haplotypes, other constraints can be obtained in the same way.

■

Theorem 3.1 and Theorem 3.2 allow us to reduce the number of possible independent haplotype level HW deviations to be estimated. For two SNPs scenario, the independent number of HW deviation parameters is 10-4=6.

Bounds of haplotype level Hardy-Weinberg deviation

The bounds of HW deviations can be established from the fact that genotypic frequencies are bounded below by zero and above by gene frequencies (Hernandez and Weir 1989):

$$0 \le P(h_i h_i) \le P_i \;\Rightarrow\; 0 \le P_i^2 + D_{ii} \le P_i \Rightarrow -P_i^2 \le D_{ii} \le P_i(1-P_i)$$

$$0 \le P(h_i h_j) \le \min(2P_i, 2P_j), where\ i \ne j$$

$$\Rightarrow\; 0 \le 2P_iP_j + D_{ij} \le \min(2P_i, 2P_j) \Rightarrow 0 \le 2P_iP_j + D_{ij} \le \min(2P_i, 2P_j)$$

$$\Rightarrow -2P_iP_j \le D_{ij} \le \min(2P_i(1-P_j), 2P_j(1-P_i))$$

For simulations, to generate a random selection of HW deviation set, individual bounds must be used for the first HW deviation, however, the sequential bounds must be used for subsequent HW deviation values chosen. Given population haplotype setting, to generate a HW deviation set for simulation studies, sequential bounds of HW deviations should be considered. Since genotype frequencies are bounded below by zero and above by haplotype frequencies, we can generate the first genotype frequency using the individual bound of the corresponding HW deviation. For the second genotype, if we use the individual bound of the corresponding HW deviation, the haplotype frequency based on selected genotypes may not match population

haplotype set. For example, suppose that we select first genotype frequency $P_{AB|Ab}$ as $2P_{AB}$ assuming $\min(2P_{AB}, 2P_{Ab}) = 2P_{AB}$. For second genotype $P_{AB|aB}$, if we select any value large than zero (i.e. within the individual bound), then the haplotype frequency P_{AB} from the two genotypes will be larger than the population haplotype frequency P_{AB} since $P_{AB|AB} + \frac{1}{2}\left(P_{AB|Ab} + P_{AB|aB} + P_{AB|ab}\right) > P_{AB}$. Therefore, sequential ranges of haplotype level HW deviations will be required for the constraints of HW deviations.

An illustrative example of sequential bounds for haplotype level HW deviations

For two SNPs scenario, there are four possible haplotypes $(1(AB), 2(Ab), 3(aB), 4(ab))$ assuming $P_1 \leq P_2 \leq P_3 \leq P_4$ and 10 unordered genotypes. Given that three genotypes $(P^B_{1|3}, P^B_{2|2}, P^B_{3|4})$ already have been selected according to corresponding HW deviations (denoting B as already selected genotypes before current genotype selection), then there are seven remaining genotypes.

	AB	Ab	aB	ab
AB	$P_{1\|1}$	$P_{1\|2}$	$P^B_{1\|3}$	$P_{1\|4}$
Ab		$P^B_{2\|2}$	$P_{2\|3}$	$P_{2\|4}$
aB			$P_{3\|3}$	$P^B_{3\|4}$
ab				$P_{4\|4}$

Haplotype frequency is obtained by summing genotypes related to the haplotype and one genotype in each group can be reparameterized in the following way:

$$P_1 = P_{1|1} + \frac{1}{2}\left(P_{1|2} + P^B_{1|3} + P_{1|4}\right) \Rightarrow P_{1|4} = 2P_1 - \left(2P_{1|1} + P_{1|2} + P^B_{1|3}\right) \geq 0$$

$$P_2 = P^B_{2|2} + \frac{1}{2}\left(P_{1|2} + P_{2|3} + P_{2|4}\right) \Rightarrow P_{2|4} = 2P_2 - \left(2P^B_{2|2} + P_{1|2} + P_{2|3}\right) \geq 0$$

$$P_3 = P_{3|3} + \frac{1}{2}\left(P^B_{1|3} + P_{2|3} + P^B_{3|4}\right) \Rightarrow P_{3|3} = P_3 - \frac{1}{2}\left(P^B_{1|3} + P_{2|3} + P^B_{3|4}\right) \geq 0$$

$$P_4 = P_{4|4} + \frac{1}{2}\left(P_{1|4} + P_{2|4} + P^B_{3|4}\right) \Rightarrow P_{4|4} = P_4 - \frac{1}{2}\left(P_{1|4} + P_{2|4} + P^B_{3|4}\right)$$

$$\Rightarrow P_{4|4} = P_4 - (P_1 + P_2) + \tfrac{1}{2}\left(\left(2P_{1|1} + P^B_{1|3}\right) + \left(2P^B_{2|2} + P_{2|3}\right) + 2P_{1|2} - P^B_{3|4}\right) \geq 0,$$

If the next HW deviation selection is for genotype $P_{1|1}$, then the inequalities related to $P_{1|1}$ should be set up to determine the corresponding sequential bounds of $P_{1|1}$. From the above reparameterization, there are two inequalities:

$$P_{1|1} \leq P_1 - \frac{1}{2}\left(P_{1|2} + P^B_{1|3}\right)$$

$$P_{1|1} \geq (P_1 + P_2) - P_4 - \frac{1}{2}\left(P_{1|3}^B + \left(2P_{2|2}^B + P_{2|3}\right) + 2P_{1|2} - P_{3|4}^B\right)$$

since $P_{1|1} \geq 0$, $\max\left(P_{1|1}\right) = \min(P_1, P_1 - \frac{1}{2}\left(P_{1|2} + P_{1|3}^B\right))$ and $\min\left(P_{1|1}\right) = \max\left(0, (P_1 + P_2) - P_4 - \frac{1}{2}\left(P_{1|3}^B + \left(2P_{2|2}^B + P_{2|3}\right) + 2P_{1|2} - P_{3|4}^B\right)\right)$. Moreover, D_{11} is bounded below by $[\max\left(P_{1|1}\right) - P_1^2]$, and bounded above by $[\min\left(P_{1|1}\right) - P_1^2]$. Sequential bounds for other HW deviations can be obtained similarly.

3.3. Analytical calculation of genotype frequency at the expectation step

For two SNPs scenarios, we denote the unphased double heterozygous (DH) genotypes frequency as $P_{DH} = P_{AB|ab} + P_{Ab|aB}$. The expectation step for $P_{AB|ab}$ at g^{th} iteration can be derived as

$$P_{AB|ab}^{(g)} = P_{DH} \frac{2P_{AB}^{(g-1)}P_{ab}^{(g-1)}}{2P_{AB}^{(g-1)}P_{ab}^{(g-1)} + 2P_{Ab}^{(g-1)}P_{aB}^{(g-1)}},$$

where $P_i^{(g-1)}s, where\ i \in (AB, Ab, aB, ab)$, are haplotype frequency estimates at the $(g-1)^{th}$ iteration.

The products of haplotypes are replaced by the corresponding sum of genotypes:

$$2P_{AB}^{(g-1)}P_{ab}^{(g-1)} = 2\left[P_{AB|AB} + \frac{1}{2}\left(P_{AB|Ab} + P_{AB|aB} + P_{AB|ab}^{(g-1)}\right)\right]\left[P_{ab|ab} + \frac{1}{2}\left(P_{AB|ab}^{(g-1)} + P_{Ab|ab} + P_{aB|ab}\right)\right]$$

$$2P_{Ab}^{(g-1)}P_{aB}^{(g-1)} = 2\left[P_{Ab|Ab} + \frac{1}{2}\left(P_{AB|Ab} + P_{Ab|aB}^{(g-1)} + P_{Ab|aB}\right)\right]\left[P_{aB|aB} + \frac{1}{2}\left(P_{AB|aB} + P_{Ab|aB}^{(g-1)} + P_{aB|ab}\right)\right].$$

When $P_{AB|ab}^{(g)} \to \hat{p}_{AB|ab}$ as $g \to \infty$ and $\hat{p}_{Ab|aB} = P_{DH} - \hat{p}_{AB|ab}$, then the above expectation step becomes a third order polynomial of $\hat{p}_{AB|ab}$ and all coefficients are function of homozygous or single heterozygous genotypes (Mano et al. 2004).

To present the coefficients, we denote $P_{AB|AB} + \frac{1}{2}\left(P_{AB|Ab} + P_{AB|aB}\right)$ as k_{AB}, $P_{ab|ab} + \frac{1}{2}\left(P_{Ab|ab} + P_{aB|ab}\right)$ as k_{ab}, $P_{Ab|Ab} + \frac{1}{2}\left(P_{AB|Ab} + P_{Ab|aB}\right)$ as k_{Ab} and $P_{aB|aB} + \frac{1}{2}\left(P_{AB|aB} + P_{aB|ab}\right)$ as k_{aB}. Then the third order polynomial of $\hat{p}_{AB|ab}$ is

$$2\hat{p}_{AB|ab}{}^3 + \{2(k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}) - 3P_{\mathrm{DH}}\}\hat{p}_{AB|ab}{}^2 + \{4(k_{\mathrm{ab}}k_{\mathrm{AB}} + k_{\mathrm{aB}}k_{\mathrm{Ab}}) - 2(k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}})P_{\mathrm{DH}} + P_{\mathrm{DH}}^2\}\hat{p}_{AB|ab} - 4k_{\mathrm{AB}}k_{\mathrm{ab}}P_{\mathrm{DH}} = 0. \quad (3.3.1)$$

The roots for $\hat{p}_{AB|ab}$ can be obtained by the cubic formula:

Root 1: $\hat{p}_{AB|ab} = -\frac{1}{3}\{2(k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}) - 3P_{\mathrm{DH}}\} + (\mathrm{S} + \mathrm{T})$ (3.3.2)

Root 2: $\hat{p}_{AB|ab} = -\frac{1}{3}\{2(k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}) - 3P_{\mathrm{DH}}\} - \frac{1}{2}(\mathrm{S} + \mathrm{T}) + \frac{1}{2}i\sqrt{3}(\mathrm{S} - \mathrm{T})$ (3.3.3)

Root 3: $\hat{p}_{AB|ab} = -\frac{1}{3}\{2(k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}) - 3P_{\mathrm{DH}}\} - \frac{1}{2}(\mathrm{S} + \mathrm{T}) - \frac{1}{2}i\sqrt{3}(\mathrm{S} - \mathrm{T})$, (3.3.4)

where S and T are defined as

$$Q \equiv \frac{1}{9}(3a_1 - a_2^2), \mathrm{R} \equiv \frac{1}{54}(-27a_0 + 9a_1a_2 - 2a_2^3), \mathrm{D} \equiv \mathrm{Q}^3 + \mathrm{R}^2$$

$$S \equiv \sqrt[3]{R + \sqrt{D}}, T \equiv \sqrt[3]{R - \sqrt{D}},$$

where $a_2 = \frac{2(k_{AB} + k_{ab} - k_{Ab} - k_{aB}) - 3P_{DH}}{2}$, $a_1 = \frac{4(k_{ab}\, k_{AB} + k_{aB}\, k_{Ab}) - 2(k_{AB} + k_{ab} - k_{Ab} - k_{aB})P_{DH} + P_{DH}^2}{2}$ and $a_0 = \frac{-4k_{AB}\, k_{ab}\, P_{DH}}{2}$.

Determining which root is real and which is complex can be categorized by the polynomial discriminant (D). If D > 0, then the equation has three distinct real roots and if D < 0, then the equation has one real root and two non real complex conjugate roots (Abramowitz and Stegun 1964). It is interesting that above roots all include a common term $[k_{AB} + k_{ab} - k_{Ab} - k_{aB}]$, which is equivalent to the difference between four homozygous genotypes as $[P_{AB|AB} + P_{ab|ab} - P_{Ab|Ab} - P_{aB|aB}]$.

3.4. Residuals from EM based haplotype frequency estimation

We will call the EM estimation error (the difference between genotype frequency and the product of constituent haplotypes) as EM residual. EM-based haplotype frequency estimation shows no residuals when phased genotype frequencies are correct. When genotype sample is in HWE, multiple heterozygous genotypes will have correct phased genotype configurations. Even when genotype sample is not in HWE, it is possible to have no EM residuals as long as phased genotype frequencies are same as true genotypes. For example, when there is no double heterozygous genotype, we can obtain haplotype frequencies without phasing process. We denote these cases when no EM residuals exist as balanced genotype settings.

When a genotype sample is unbalanced, the products of haplotype frequency estimations are no longer the same as genotype frequencies, indicating that the genotype sample is deviated from HWE. However, EM residuals are not the true haplotype level HW deviations.

3.5. Simulation settings

To sample genotypes from the population, Fallin and Schork (2000) assigned a first haplotype to each individual, with probabilities equal to the population haplotype frequencies. Then they used conditional probabilities for each haplotypes to assign the second haplotype for each individual and the joint probabilities could be expressed as functions of SNP level HW deviations. However, haplotype level HW deviations are not fully investigated. The primary goal of our simulation is to explore the haplotype level HW deviations from a particular haplotype setting. The simulation settings are summarized in Table 1.

Population settings

For two SNPs scenario, population settings for HW deviations and genotypes are specified for different haplotype scenarios. Given population haplotype frequencies, genotype frequencies under HWE is simply product of two constituent haplotypes. Haplotype level HW deviations can be randomly generated by predetermined sequential HW deviation bounds. For two SNPs scenario, 100,000 HW deviation sets were selected each from equal haplotype frequencies setting (0.25, 0.25, 0.25, 0.25) and unequal haplotype frequencies setting (0.1, 0.2, 0.3, 0.4). Population genotype frequency can be obtained from the selected HW deviation

scenarios of a particular haplotype setting by adding to the product of two constituent haplotype frequencies.

Sampling setting

Genotype samples with sample size n are randomly chosen from each population genotype frequency setting. The true sample haplotype frequencies are calculated by counting the number of occurrences of each haplotype in the genotype sample and divided by the total number of haplotypes (2n). As double heterozygous genotypes play the critical role in haplotype determination, for equal or unequal haplotype frequency settings, we sort the simulated data according to the sum of double heterozygous genotype frequencies. We choose the following percentiles for further investigation (1^{st}, 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} and 99^{th}) from simulated data. To capture reasonable number of genotype sets at each percentile, we pick a small bin by adding or subtracting 0.0001 from each chosen percentile of sum of double heterozygous genotype frequencies (e.g. at 50^{th} percentile, we use the bin from [50^{th} percentile - 0.0001, 50^{th} percentile + 0.0001]). Population genotype sets within each bin are selected to start sampling. Samples are repeated 200 times with different sample sizes (25, 50, 100 and 500) and means (standard deviations) of sampling error (MSE_{PS}) are calculated. Double heterozygous genotypes from the samples are then unphased based on EM algorithm to investigate the haplotype frequency estimation error. The means and standard deviations for estimation errors (MSE_{SE} and MSE_{PE}) are also obtained. Four mean squared errors are specified to distinguish the source of estimation error: 1). EM estimation error between population haplotype and EM estimated haplotype for the population genotype setting (MSE_{PEp}), 2). EM estimation error between sample haplotype and sample estimated haplotype frequency (MSE_{SEs}), 3). sampling

error between population haplotype and sample haplotype frequency from genotype sample (MSE_{PS}) and 4). error between population haplotype and estimated haplotype frequency from genotype sample (MSE_{PEs}). To avoid local maxima issue of EM algorithm, different initial values are used and we set the convergence criteria of 10^{-6} based on absolute change in log likelihood similar to that of program haplo.em in R package.

3.6. Simulation results

Analytical calculation of genotype frequency on the expectation step

We plot the difference between true and estimated frequency for one of the double heterozygous genotype ($P_{AB|ab} - \hat{p}_{AB|ab}$) against $[k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}]$ based on 10,000 random HW deviation scenarios from equal and unequal haplotype frequency setting (Figures A1.1 and A1.2). There exist distinct areas according to the cubic discriminant (D) and the root identification from (3.3.2) or (3.3.3). The root (3.3.4) is not shown since the value is out of genotype boundary. For equal haplotype frequency setting, the theoretical maximum of the difference ($P_{AB|ab} - \hat{p}_{AB|ab}$) is 0.5 when $P_{AB|ab}$ is its maximum as $\min(2P_{AB}, 2P_{ab}) = 0.5$ from the genotype bounds and $\hat{p}_{AB|ab} = 0$. When $[k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}]$ is deviated from zero, the genotype has severe estimation error that would result in incorrect haplotype frequency estimations.

Genotype frequency estimation and HW deviations for double heterozygous genotypes

A root for double heterozygous genotype estimation is shown to be a function of the unphased genotype frequencies in section 3.3. The relationship between the root and HW deviations for double heterozygous genotypes fixing haplotypes can also be explored by simulations. We can parameterize all coefficients of the third order polynomial in terms of haplotype frequencies and HW deviations of double heterozygous genotypes as:

$$k_{\mathrm{AB}} = P_{AB|AB} + \frac{1}{2}\left(P_{AB|Ab} + P_{AB|aB}\right) = P_{AB}^2 + D_{AB|AB} + \frac{1}{2}\left(2P_{AB}P_{Ab} + D_{AB|Ab} + 2P_{AB}P_{aB} + D_{AB|aB}\right)$$

$$= P_{AB}^2 + P_{AB}P_{Ab} + P_{AB}P_{aB} + D_{AB|AB} + \frac{1}{2}\left(D_{AB|Ab} + D_{AB|aB}\right) = P_{AB}(1 - P_{ab}) - \frac{1}{2}D_{AB|ab}$$

$$k_{\mathrm{ab}} = P_{ab}(1 - P_{AB}) - \frac{1}{2}D_{AB|ab},\ k_{\mathrm{Ab}} = P_{Ab}(1 - P_{aB}) - \frac{1}{2}D_{Ab|aB},\ k_{\mathrm{aB}} = P_{aB}(1 - P_{Ab}) - \frac{1}{2}D_{Ab|aB}$$

$$P_{\mathrm{DH}} = 2P_{AB}P_{ab} + D_{AB|ab} + 2P_{Ab}P_{aB} + D_{Ab|aB}.$$

Given that the sum of the two unphased double heterozygous genotype frequencies (P_{DH}) is fixed and haplotype frequencies are known, then all coefficients of the third order polynomial are functions of haplotypes and the HW deviations corresponding to the root which is one of double heterozygous genotypes. It is challenging to figure out the relationship between the genotype frequency estimation and the corresponding HW deviation analytically since the root itself has complicate form. However, we can investigate the relationship through simulation.

Since the true genotype frequency is determined by the sum of product of haplotype frequencies and corresponding haplotype level HW deviation, the amount of the genotypes increases as the corresponding HW deviation increases when haplotype frequencies are fixed. However, the genotype frequency decreases when the true genotype frequency increases under different levels of P_{DH} for equal and unequal haplotype frequency setting (Figure A1.3 and A1.4). This result also shows that the haplotype frequency estimation based on EM algorithm under HWE assumption could result in incorrect genotype frequency estimation at expectation step because of HW deviation.

EM estimation error between population haplotype and EM estimated haplotype for the population genotype setting

For equal haplotype frequency setting, the relationship between sum of double heterozygous genotypes in population and EM estimation error for population genotype (MSE_{PEp}) is illustrated (Figure A1.5). Theoretical maximum MSE_{PEp} is 0.0625 when two estimated haplotype frequencies are 0.5 and the other two are zero, i.e. $\max(MSE_{PEp}) = \frac{(0.25-0.5)^2+(0.25-0)^2+(0.25-0)^2+(0.25-0.5)^2}{4} = 0.0625$. This error comes from incorrect estimation for double heterozygous genotype assuming HWE. Since one of double heterozygous genotype frequency is bounded above by 0.5, when the P_{DH} is larger than 0.5, MSE_{PEp} could reach up to its theoretical maximum. For example, when the P_{DH} is 0.6 and $k_{AB} = k_{aB} = 0, k_{ab} = k_{Ab} = 0.2, P_{AB|ab} = 0.5 \; and \; P_{Ab|aB} = 0.1$, then population haplotype frequency is $P_{AB} = k_{AB} +$

$\frac{1}{2}P_{AB|ab} = 0.25$, $P_{aB} = k_{aB} + \frac{1}{2}P_{Ab|aB} = 0.25$, $P_{Ab} = k_{Ab} + \frac{1}{2}P_{Ab|aB} = 0.25$, and $P_{ab} = k_{ab} + \frac{1}{2}P_{AB|ab} = 0.25$. The maximum estimation error occurs when EM estimated haplotype frequency is $\hat{p}_{AB} = \hat{p}_{ab} = 0\ , and\ \hat{p}_{Ab} = \hat{p}_{aB} = 0.5$ from estimated genotype frequency $\hat{p}_{AB|ab} = 0$ and $\hat{p}_{Ab|aB} = 0.6$. When P_{DH} is less than 0.5, MSE_{PEp} is bounded above because estimated haplotype frequency is always less than 0.5, and MSE_{PEp} can be zero when EM estimated haplotype is evenly distributed, i.e. balanced cases.

Unequal haplotype frequency setting has a more complicated pattern because of unequal haplotype frequency (Figure A1.9). Since there is a difference between the double heterozygous genotype frequencies in HWE, the bounds of the genotypes are also different and the interval of EM estimated genotype frequency is from zero to P_{DH}. This means that the absolute differences of $|P_{AB|ab} - \hat{p}_{AB|ab}|$ and $|P_{Ab|aB} - \hat{p}_{Ab|aB}|$ have different bounds resulting for different levels of MSE_{PEp}.

Sampling error vs. estimation error

Comparing MSE_{PEs} and MSE_{PS} with MSE_{SEs} at different levels of MSE_{PEp} in our simulation study showed different results and it depends on the sum of double heterozygous genotypes (P_{DH}). When P_{DH} is low (1st, 5th, 10th and 25th percentiles), sampling error is dominant and similar to Fallin and Schork (2000) have found (Tables A2.1 and A2.2, Figures A1.6 and A1.10). However, when P_{DH} is relatively high, estimation error is at least comparable to sampling error (75, 90, 95 and 99 percentiles) (Figures A1.6 and A1.10). We find that MSE_{SEs}

can be more serious than MSE_{PS} at even small sample size (25) for larger than 50th percentile of P_{DH} for equal and unequal haplotype settings (Tables A2.1 and A2.2). The pattern of severe estimation error is clearer when we fix the sample size at 100 for equal and unequal haplotype settings (Figures A1.7 and A1.11). When MSE_{PEp} increases, MSE_{SEs} and MSE_{PEs} also increase whereas MSE_{PS} is consistent even though the standard deviations of MSE_{PEs} and MSE_{SEs} are larger than the standard deviations of MSE_{PS}. The sampling error is shown to converge to zero as the sample size increases. The estimation error MSE_{PEs} converges to the true estimation error MSE_{PEp} with increasing sample size. One should be careful to estimate haplotype frequencies via EM algorithm when the true genotypes are deviated from HWE.

The sum of double heterozygous genotype frequencies can be a measure of missing phase information since single or homozygous genotypes can be phased without error. It is found that the estimation error depends on the sum of double heterozygous genotypes (Figures A1.5 and A1.9). When the sum of double heterozygous genotype frequencies is fixed, the estimation error varies according to the actual genotype setting. Heterozygosity might not reflect the information because it includes single heterozygous genotype frequencies which can be phased without any estimation process. It is obvious that the genotype setting with excessive homozygosity as close to one has relatively small estimation error since it has small amount of double heterozygous genotypes. However, it doesn't guarantee that increased homozygosity (same as decreased heterozygosity) has always decreased estimation error and increased heterozygosity (same as decreased homozygosity) has always increased estimation error (Figures A1.8 and A1.12).

Chapter 4

HWD-ECM: a proposed method for haplotype and HW deviation joint estimation

4.1. HWD-ECM approach for haplotype frequency and HW deviation parameter estimation

We studied the impact of HW deviation on EM-based haplotype frequency estimation in chapter 3. In this chapter, we propose a new method which incorporates HW deviation parameters into the haplotype frequency estimation process. The proposed Hardy-Weinberg Deviation-Expectation/Conditional Maximization (HWD-ECM) method enables to estimate haplotype frequencies and HW deviations simultaneously.

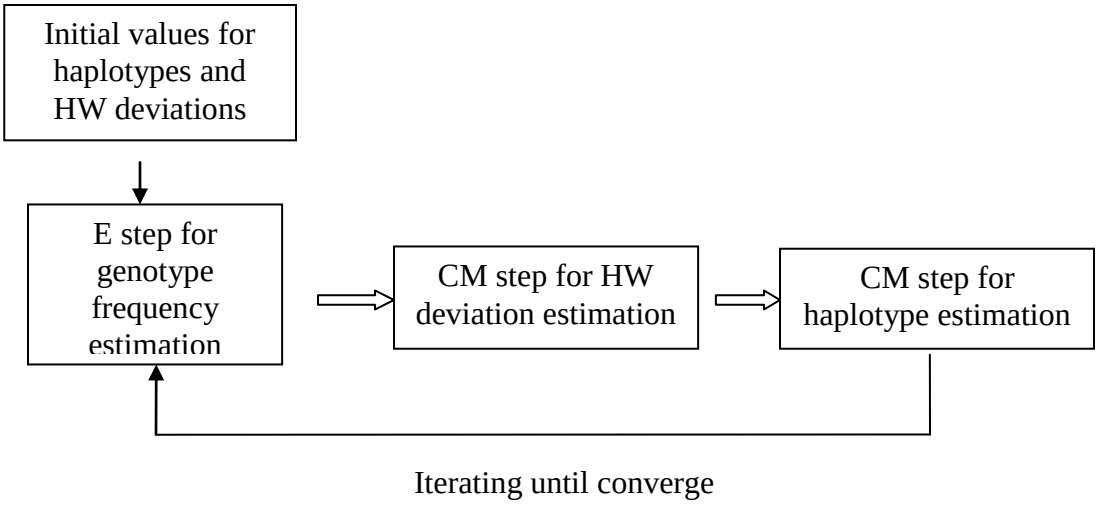
When the genotype frequencies are not in HWE, HW deviation parameters have to be incorporated for genotype frequencies and the complete log likelihood is modified:

$$\log(L_c) = \sum_{(k,l)\in h} n_{kl} \log(P(h_k h_l))$$

$$= \sum_{t=1}^{h} \{ n_{tt} \log(P_t^2 + d_{tt}) + \sum_{i=1, i\neq t} n_{ti} \log(2P_t P_i + d_{ti}) \}.$$

The traditional EM approach does not allow HW deviation parameters to be estimated. It is also not attractive to use the EM algorithm when adding HW deviation parameters since its maximization step will be too complicated. The HWD-ECM approach can resolve this problem

by using two conditional maximization steps rather than a single maximization step. The following diagram summarizes the whole process.



4.1.1. Expectation step allowing for HW deviations

The expectation step of the HWD-ECM approach is constructed by modifying the expectation step of EM-based haplotype frequency estimation adding HW deviation parameters:

$$\hat{P}(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{P(genotype\ h_k h_l\ in\ phenotype\ j)}{P(phenotype\ j)} = \frac{n_j}{n} \frac{P_j(h_k h_l)^{(g)}}{P_j{}^{(g)}},$$

where $P_j(h_k h_l)^{(g)} = \begin{cases} (p_k^{(g)})^2 + d_{kk}^{(g)}, & if\ k = l, \\ 2p_k^{(g)} p_l^{(g)} + d_{kl}^{(g)}, & if\ k = l. \end{cases}$

Like EM-based haplotype frequency estimation, the initial values of haplotype frequency and HW deviations will be needed to start the HWD-ECM algorithm. Even though the initial haplotypes can be randomly chosen, the HW deviations should be selected within the boundaries

from the initial haplotypes considering the constraints of HW deviations. For simplicity, zero HW deviations (HWE) are used as initial haplotype frequencies.

4.1.2. Conditional maximization step for HW deviation parameter estimation

To obtain the maximum likelihood estimation of HW deviation parameters, score functions which are partial derivatives with respect to each HW deviation parameter are used conditioning on other parameters. Since the sum of HW deviation parameter values should be zero, it is necessary to solve the differential equation with this constraint. Lagrange multiplier λ_{HWD} can be used to find the maximum likelihood estimates of HW deviation parameters subject to the constraint. When we take a partial derivative with respect to d_{ij},

$$\frac{\partial \log(L_c)}{\partial d_{ij}} + \lambda_{HWD} \Longrightarrow \begin{cases} \dfrac{n_{ii}}{p_{ii}^2 + d_{ii}} + \lambda_{HWD} = 0, \;\; if\ i = j, \\ \dfrac{n_{ij}}{2p_i p_j + d_{ij}} + \lambda_{HWD} = 0, \;\; if\ i \neq j \end{cases}$$

Since $\sum_{i=1}^{h} \sum_{j=1, i \neq j}^{h} d_{ij} = 0$, the Lagrange multiplier λ_{HWD} can be analytically obtained by simple algebraic procedures as –n which is negative of total number of individuals in the sample. The (g+1) th iteration of HW deviation estimates conditioning on the above Lagrange multiplier and the g th iteration of haplotype frequency estimates can be written as

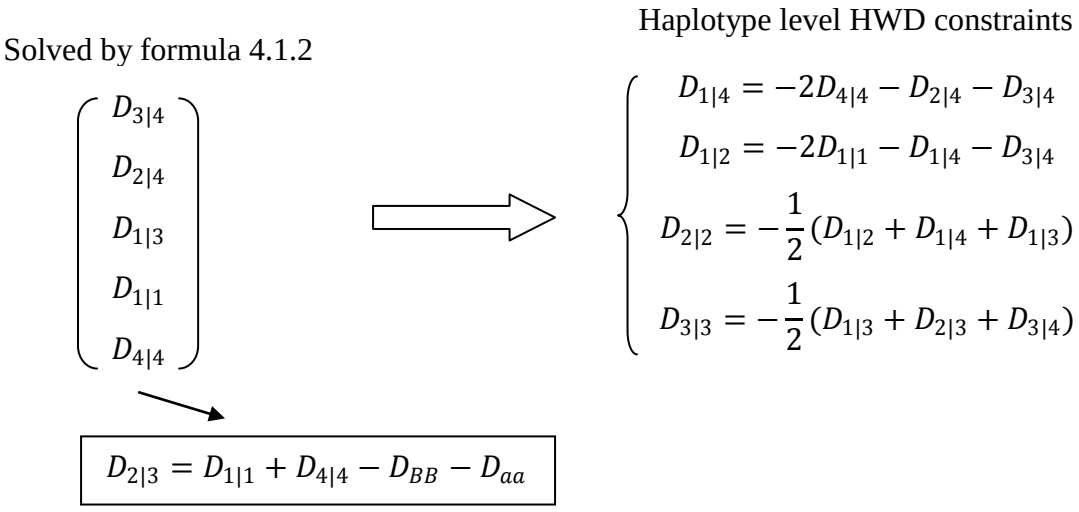
$$d_{ii}^{(g+1)} = \frac{n_{ii}}{n} - (p_i^{(g)})^2, if\ i = j,$$

$$d_{ij}^{(g+1)} = \frac{n_{ij}}{n} - 2p_i^{(g)} p_j^{(g)}, if\ i \neq j. \qquad (4.1.2)$$

Besides the Lagrange multiplier λ_{HWD}, the additional constraints of HW deviation related to each haplotype should be considered. For two SNPs scenarios, based on the constraints of haplotype level HW deviations, we can reduce the four HW deviation parameters in terms of six haplotype level HW deviations obtained by (4.1.2). For example, four HW deviations including one of double heterozygous genotypes can be calculated by remaining six HW deviations. It is important to determine which four should be selected out of ten HW deviation parameters. Since the constraints can be treated as equations of four unknown variables (HW deviations), one HW deviation parameter of each equation can be solved in terms of other parameters. Since HW deviation parameters of single or double heterozygous genotypes exist simultaneously in two different equations, particular settings of four HW deviations need be chosen. We select $D_{1|2}, D_{1|4}, D_{2|2}\ and\ D_{3|3}$, where $1: AB, 2: Ab, 3: aB, and\ 4: ab$. Base on haplotype level HW deviation constraints, we can substitute $D_{1|2}, D_{1|4}, D_{2|2}\ and\ D_{3|3}$ in terms of other six HW deviations:

$$D_{1|4} = -2D_{4|4} - D_{2|4} - D_{3|4}$$
$$D_{1|2} = -2D_{1|1} - D_{1|4} - D_{3|4}$$
$$D_{2|2} = -\frac{1}{2}(D_{1|2} + D_{1|4} + D_{1|3})$$
$$D_{3|3} = -\frac{1}{2}(D_{1|3} + D_{2|3} + D_{3|4})$$

In addition to the constraint of haplotype level HW deviation, we can exploit SNP level HW deviation information based on Theorem 3.1. We substitute one double heterozygous genotype using SNP level HW deviations and other haplotype level HW deviations. To summarize the process for two SNPs scenario, we illustrated this CM step using a simple diagram labeling haplotypes AB, aB, Ab and ab as 1, 2, 3 and 4 respectively:



4.1.3. Conditional maximization step for haplotype frequency estimation

The conditional maximization step for t [th] haplotype frequency of (g+1) [th] iteration was obtained in a similar way as the maximization step of EM-based haplotype frequency estimation:

$$p_t{}^{(g+1)} = P(h_t h_t)^{(g)} + \frac{1}{2}\sum_{i=1, i\neq t}^{h} P(h_t h_i)^{(g)}.$$

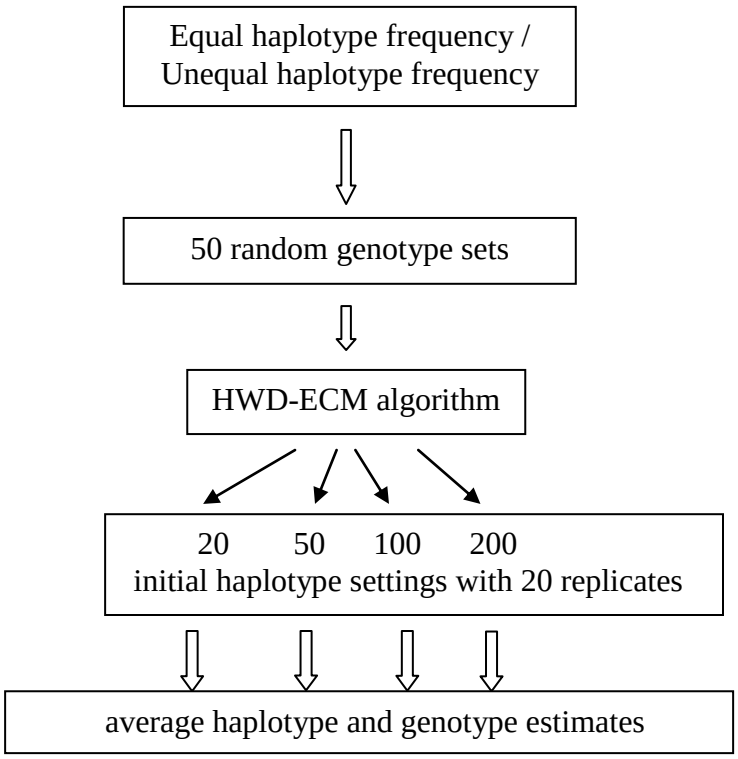
This process updates genotype frequency by the product of haplotype frequency and corresponding HW deviation estimates from the previous iteration. Since the configurations of haplotypes in single or homozygous genotypes are already determined, only multiple heterozygous genotypes need to be updated by the expectation step (e.g. double heterozygous genotypes for two SNPs scenarios).

4.2. Simulation settings

One technical issue of allowing HW deviations at the expectation step is that it is possible to have the sum of haplotype frequency product and HW deviation being negative at certain iteration. It can cause incorrect phasing for multiple heterozygous genotypes since negative genotype frequency is not allowed. The problem can be avoided by forcing the negative values to be zero. The estimated genotype frequencies are then sent to the conditional maximization steps.

Based on one expectation step and two conditional maximization steps, when we insert one initial haplotype frequency set and zero HW deviations, it converges under the criteria of 10^{-6} based on absolute change in log likelihood. To investigate the convergence on different initial haplotype frequency, 5000 random initial haplotypes were used in the HWD-ECM algorithm for 50 randomly selected genotype settings each for equal and unequal haplotype frequency settings. The means (standard deviations) of MSE_{PEp}s from HWD-ECM method were compared with MSE_{PEp} based on EM method.

To investigate the performance of different numbers of initial haplotypes for HWD-ECM, 25, 50, 100 and 200 random initial haplotypes were studied with 20 replicates. For 50 random genotype sets each from equal and unequal haplotype setting, we plot MSE_{PEp}s to check the performance of HWD-ECM compared with EM method. We illustrate the process by following diagram.



To further illustrate the possible improvement of our method, we select five genotype settings where MSE_{PEp}>0.01 based on EM algorithm for each equal and unequal haplotype frequency setting. Sampling procedure is similar to that in section 3.5. For each of the five genotypes, samples are repeated 50 times with different sample sizes (50, 100 and 500) and estimation errors (MSE_{SEs}) from HWD-ECM algorithm are compared to estimation errors from EM algorithm.

4.3. Simulation results

Based on MSE_{PEp} from 50 randomly selected genotype sets of equal and unequal haplotype frequency setting, HWD-ECM algorithm performs better than EM algorithm (Figure

A1.13 and A1.15). We obtain averages of 5000 MSE_{PEp}s with standard deviations according to 5000 initial haplotypes. For equal haplotype frequency setting, HWD-ECM performs significantly better than EM algorithm where $MSE_{PEp} > 0.01$. In average, 0.01 level of MSE represents absolute difference of true and estimated haplotype of 0.1 based on MSE calculation. For low level of MSE_{PEp} (<0.0001), the absolute difference of true and estimated haplotype is about 0.01. HWD-ECM may perform worse than EM algorithm for low level of MSE_{PEp}. However, the absolute difference might be acceptable. For example, MSE_{PEp} of EM algorithm is 0.0000001 and MSE_{PEp} of HWD-ECM is 0.0001. The absolute difference is 0.0001 from the EM algorithm and 0.01 from HWD-ECM algorithm even though EM algorithm performed 10000 times better than HWD-ECM in term of MSE. Therefore, we should have more attention to high level of MSE_{PEp}. The performance of HWD-ECM for unequal haplotype frequency setting is similar.

There is no large difference among 25, 50 and 100 initial numbers (Figure A1.14 and A1.16). As EM MSE_{PEp} increases, HWD-ECM performs well for different number of initial haplotypes. We use 100 initial haplotypes for HWD-ECM to make sure stable estimates. For 750 samples (5 genotype setting × 3 different sample size × 50 replicates) each from equal and unequal haplotype settings, we confirm that HWD-ECM performs better than EM algorithm using two ratios of estimation errors (EM estimation error/ HWD-EDM estimation error, HWD-ECM estimation error/ EM estimation error) (Figure A1.17 and A1.18). Four cases from HWD-ECM are worse than EM algorithm for equal haplotype frequency setting and a single case for unequal haplotype frequency setting. Since the levels of estimation error are relatively low, the performances of EM and HWD-ECM are similar.

4.4. Extension to multiple SNPs scenarios

The HWD-ECM approach for two SNPs scenarios can be extended to multiple SNPs scenarios. It requires sequential steps to estimate multiple heterozygous genotype frequencies. As the current HWD-ECM approach for two SNPs scenarios can estimate two double heterozygous genotypes, conditioning on alleles of other SNPs enables us to estimate each two double heterozygous genotypes by constructing multiple two SNPs sets within a multiple SNPs scenarios.

When the number of SNPs is k, the number of haplotypes is 2^k. The genotype can be summarized as either ordered genotypes which consider the order of two constituent haplotypes or the unordered genotypes without considering the order, so the total number of the ordered genotypes is 2^{2k} and the total number of the unordered genotypes is $\frac{2^k(2^k+1)}{2}$. From now on, we consider the unordered genotypes only. Furthermore, the number of homozygous genotypes is same as the number of haplotypes (2^k) and the number of unordered heterozygous genotypes is $\frac{2^k(2^k+1)}{2} - 2^k = \frac{2^k(2^k-1)}{2}$. The number of m-tuple unordered heterozygous genotypes is $\binom{k}{m}2^{k-1}$, where $0 < m \leq k$.

An illustrative example: three SNPs scenarios

We denote H as homozygous genotype, S as single heterozygous genotype, D as double heterozygous genotype and T as triple heterozygous genotype. For a three SNPs scenario, we have 8 Hs, 12 Ss, 12 Ds and 4 Ts. All possible genotype features of three SNPs scenarios are illustrated (Figure A1.19). When we relabel haplotype "ABC", "aBC", "AbC", "abC", "ABc",

"aBc", "Abc" and "abc" as 1 to 8, each genotype can be identified as one pair of the numbers, e.g. ABC|AbC by S13.

By conditioning on a particular allele for each SNP one by one, there are six sets of two SNPs scenarios (Figure A1.20). When the HWD-ECM approach is applied to each set, all double heterozygous genotypes can be phased. The sum of all genotype frequencies of each set can then be multiplied by the estimated double heterozygous genotype frequencies to obtain joint probability.

To derive the triple heterozygous genotype frequencies, we merge the genotypes according to each SNP. There are three sets of two SNPs scenarios (Figure A1.21). For the first SNP, unphased genotype frequency is

$P_{BC|bc} + P_{Bc|bC} = D17 + T18 + T27 + D28 + D35 + T36 + T45 + D46$ by merging "A" or "a" for first SNP. The unphased genotype can be phased into

$$P_{BC|bc} = D17 + T18 + T27 + D28,\ P_{Bc|bC} = D35 + T36 + T45 + D46$$

For the second SNP,

$P_{AC|ac} + P_{Ac|aC} = D16 + T18 + T36 + D38 + D25 + T27 + T45 + D47$ by merging "B" or "b" can be phased into

$$P_{AC|ac} = D16 + T18 + T36 + D38,\ P_{Ac|aC} = D25 + T27 + T45 + D47$$

For the third SNP,

$P_{AB|ab} + P_{Ab|aB} = D14 + T18 + T45 + D58 + D23 + T27 + T36 + D67$ by merging "C" or "c" can be phased into

$$P_{AB|ab} = D14 + T18 + T45 + D58,\ P_{Ab|aB} = D23 + T27 + T36 + D67$$

Then we can set up a system of equations for triple heterozygous genotypes:

$$T18 + T27 = P_{BC|bc} - (D17 + D28),$$

$$T36 + T45 = P_{Bc|bC} - (D36 + D46),$$

$$T18 + T36 = P_{AC|ac} - (D16 + D38),$$

$$T27 + T45 = P_{Ac|aC} - (D25 + D47),$$

$$T18 + T45 = P_{AB|ab} - (D14 + D58),$$

$$T27 + T36 = P_{Ab|aB} - (D23 + D67),$$

and T18+T27+T36+T45=P(sum of triple heterozygous genotypes).

We can obtain phased triple heterozygous genotype frequencies by solving the above system of equations. For three SNPs scenario, we need to run six HWD-ECM algorithm runs for double heterozygous genotypes and three HWD-ECM runs for triple heterozygous genotypes. In total, we need nine runs of HWD-ECM algorithm to solve a three SNPs scenario.

Chapter 5

Discussion

The Expectation Maximization (EM) algorithm is widely used in haplotype phasing and frequency estimation. When Hardy-Weinberg (HW) equilibrium assumption is violated, the estimation errors for haplotype and genotype frequency can be severe. This estimation error also impacts the downstream genetic analysis, e.g. case-control risk analysis based on estimated haplotypes.

Our simulations show that the EM algorithm can result in more severe estimation error than sampling error for haplotypes and genotypes. With increasing sample size, the estimation error (MSE_{SE}) converges to its true estimation error (MSE_{PEp}) because of the existence of haplotype level HW deviations. The increased homozygosity (same as decreased heterozygosity) does not guarantee decreased estimation error and increased heterozygosity (same as decreased homozygosity) does not always increase estimation error. The estimation errors are related to incorrect estimation of multiple heterozygous genotypes frequencies (double heterozygous genotypes in two SNPs cases) since the EM expectation step assumes HWE.

SNP level HW deviations are not very useful for the investigation of multiple loci genotype frequency estimation space because the variation of genotype frequencies is directly produced by haplotype level HW deviations. Based on Theorem 3.1, haplotype level HW deviation can be severe when all SNP level HW deviations are zero (Corollary 3.1 and 3.2).

Therefore, haplotype frequency estimation error should be explored in terms of haplotype level HW deviation and cover the full range of genotype frequency space.

To modify the incorrect HWE assumption in EM algorithm, we developed a HWD-ECM algorithm to estimate haplotype frequencies as well as HW deviation parameters. Simulation results show that the HWD-ECM method performs significantly better than the EM-based approach in haplotype estimation when HWE assumption is violated.

Limitations and future work

Ideally, we would want a single execution of HWD-ECM algorithm to converge to true haplotype frequencies and HW deviations. However, because of the identifiable issue of our parameterization of HW deviations and haplotype frequencies, the results depend on the initial haplotypes. By incorporating SNP level HW deviation, we narrow down the parameter space and improve the estimation. We provide a workable solution by averaging haplotype frequency estimates from multiple initial haplotype values. The results are significantly better than those from EM algorithm. There are several other limitations on the HWD-ECM algorithm.

First, the computational issue of running HWD-ECM appears with multiple initials and multiple SNPs scenario. The more initial haplotypes, the more computing will be needed. There is no dramatic difference among 25, 50, 100 and 200 initials (Figure A1.14, A1.16). We recommend using 100 initial haplotypes to make sure the results are stable. The computational burden can also be severe for multiple SNPs scenarios. Even for three SNPs scenario, we need to run HWD-ECM algorithm nine times. For multiple SNPs scenario, one way to relieve computational burden is to develop a fast algorithm such as Partition-Ligation approach (Qin et

al. 2002). If we can obtain partitioned haplotype frequencies and HW deviations by parallel calculations, then much of computational time would be saved.

Second, we need to consider additional information (constraint) about haplotype level HW deviations. We did not obtain unique solution set which are true haplotypes and HW deviations via Theorem 3.1 and the constraints of SNP and haplotype level HW deviations. In other words, there are too many parameters to be estimated. When we fix single haplotype level HW deviation value, the algorithm has no issue of initial haplotypes and unique solution. Furthermore, when we fix the haplotype level HW deviation at true value, it converges to true haplotypes and HW deviations with zero estimation error. It is worthwhile to explore additional information of SNP level or haplotype level HW deviation. One approach is to develop a model selection procedure, step-be-step, reducing the number of parameters needs to be established.

Third, we did not investigate analytic solutions of the differential equation. We obtained four HW deviation parameters by calculating the other six HW deviations rather than by substituting four HW deviation parameters in terms of other six HW deviations in the differential equation. There are advantages of using substitution as it does not require incorporation of the Lagrange multiplier λ_{HWD}. However, one possible issue of the substitution approach is that multiple roots can occur in the process of solving the differential equations. Even though the multiple roots issue can be solved by several methods (Barnett 1966; Small et al. 2000), they were all based on the existence of unique and consistent estimator under regularity conditions. However, those methods may not applicable because haplotype frequency and HW deviation parameter following a multinomial distribution violates the one of regularity conditions. In other words, the haplotype frequency and HW deviation parameters are non-identifiable since the

different combinations of haplotype and HW deviation can produce same genotype frequency which causes the same likelihood.

Satten and Epstein (2004) attempted to estimate haplotype frequencies, single common fixation index (F) parameter and relative risk parameters simultaneously in their case control study. However, it is not practical to use single fixation index (F). The impact of HW deviation on relative risk estimation in case control studies should be investigated. We will modify the HWD-ECM approach to allow estimation of haplotype frequency and HW deviation parameters, and relative risk parameters simultaneously for case control studies.

We did not compare with Bayesian approaches, which would be useful to guide practitioner in choosing the most appropriate haplotype estimation programs for their research.

References

Abramowitz, M. and I. A. Stegun (1964). Handbook of mathematical functions with formulas, graphs, and mathematical tables. Washington, U.S. Govt. Print. Off.

Arnheim, N., P. Calabrese and M. Nordborg (2003). "Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved." American Journal of Human Genetics **73**(1): 5-16.

Barnett, V. D. (1966). "Evaluation of Maximum-Likelihood Estimator Where Likelihood Equation Has Multiple Roots." Biometrika **53**: 151-&.

Ceppellini, R., M. Siniscalco and C. A. B. Smith (1955). "The Estimation of Gene Frequencies in a Random-Mating Population." Annals of Human Genetics **20**(2): 97-115.

Chen, J. J., T. Duan, R. Single, K. Mather and G. Thomson (2005). "Hardy-Weinberg testing of a single homozygous genotype." Genetics **170**(3): 1439-1442.

Chen, J. J. and G. Thomson (1999). "The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test." Biometrics **55**(4): 1269-1272.

Clark, A. G. (1990). "Inference of Haplotypes from Pcr-Amplified Samples of Diploid Populations." Molecular Biology and Evolution **7**(2): 111-122.

Cox, D. G. and P. Kraft (2006). "Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error." Human Heredity **61**(1): 10-14.

Daly, M. J., J. D. Rioux, S. E. Schaffner, T. J. Hudson and E. S. Lander (2001). "High-resolution haplotype structure in the human genome." Nature Genetics **29**(2): 229-232.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data Via Em Algorithm." Journal of the Royal Statistical Society Series B-Methodological **39**(1): 1-38.

Douglas, J. A., M. Boehnke, E. Gillanders, J. A. Trent and S. B. Gruber (2001). "Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies." Nature Genetics **28**(4): 361-364.

Elston, R. C. and R. Forthofer (1977). "Testing for Hardy-Weinberg Equilibrium in Small Samples." Biometrics **33**(3): 536-542.

Emigh, T. H. (1980). "A Comparison of Tests for Hardy-Weinberg Equilibrium." Biometrics **36**(4): 627-642.

Epstein, M. P. and G. A. Satten. (2003, Dec). "Inference on haplotype effects in case-control studies using unphased genotype data." American Journal of Human Genetics Retrieved 6, 73, from <Go to ISI>://000187491100009

Excoffier, L. and M. Slatkin (1995). "Maximum-Likelihood-Estimation of Molecular Haplotype Frequencies in a Diploid Population." Molecular Biology and Evolution **12**(5): 921-927.

Fallin, D., A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen and N. J. Schork (2001). "Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease." Genome Research **11**(1): 143-151.

Gomes, I., A. Collins, C. Lonjou, N. S. Thomas, J. Wilkinson, M. Watson and N. Morton (1999). "Hardy-Weinberg quality control." Annals of Human Genetics **63**: 535-538.

Guo, S. W. and E. A. Thompson (1992). "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles." Biometrics **48**(2): 361-372.

Hardy, G. H. (1908). "Mendelian proportions in a mixed population." Science **28**: 49-50.
Hawley, M. E. and K. K. Kidd (1995). "Haplo - a Program Using the Em Algorithm to Estimate the Frequencies of Multisite Haplotypes." Journal of Heredity **86**(5): 409-411.
Hernandez, J. L. and B. S. Weir (1989). "A Disequilibrium Coefficient Approach to Hardy-Weinberg Testing." Biometrics **45**(1): 53-70.
Kuk, A. Y. C., H. Zhang and Y. Yang (2009). "Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium." Bioinformatics **25**(3): 379-386.
Li, S. S. Y., N. Khalid, C. Carlson and L. P. Zhao (2003). "Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms." Biostatistics **4**(4): 513-522.
Long, J. C., R. C. Williams and M. Urbanek (1995). "An E-M Algorithm and Testing Strategy for Multiple-Locus Haplotypes." American Journal of Human Genetics **56**(3): 799-810.
Mano, S., N. Yasuda, T. Katoh, K. Tounai, H. Inoko, T. Imanishi, G. Tamiya and T. Gojobori (2004). "Notes on the maximum likelihood estimation of haplotype frequencies." Annals of Human Genetics **68**: 257-264.
Meng, X. L. and D. B. Rubin (1993). "Maximum-Likelihood-Estimation Via the Ecm Algorithm - a General Framework." Biometrika **80**(2): 267-278.
MichalatosBeloin, S., S. A. Tishkoff, K. L. Bentley, K. K. Kidd and G. Ruano (1996). "Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR." Nucleic Acids Research **24**(23): 4841-4843.
Niu, T. H. (2004). "Algorithms for inferring haplotypes." Genetic Epidemiology **27**(4): 334-347.
Niu, T. H., Z. H. S. Qin, X. P. Xu and J. S. Liu (2002). "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms." American Journal of Human Genetics **70**(1): 157-169.
Qin, Z. H. S., T. H. Niu and J. S. Liu (2002). "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms." American Journal of Human Genetics **71**(5): 1242-1247.
Satten, G. A. and M. P. Epstein (2004). "Comparison of prospective and retrospective methods for haplotype inference in case-control studies." Genetic Epidemiology **27**(3): 192-201.
Schaid, D. J. (2004). "Genetic epidemiology and haplotypes." Genetic Epidemiology **27**(4): 317-320.
Small, C. G., J. F. Wang and Z. J. Yang (2000). "Eliminating multiple root problems in estimation." Statistical Science **15**(4): 313-332.
Smith, C. A. B. (1957). "Counting Methods in Genetical Statistics." Annals of Human Genetics **21**(3): 254-276.
Sobrino, B., M. Brion and A. Carracedo (2005). "SNPs in forensic genetics: a review on SNP typing methodologies." Forensic Science International **154**(2-3): 181-194.
Stephens, J. C., J. Rogers and G. Ruano (1990). "Theoretical Underpinning of the Single-Molecule-Dilution (Smd) Method of Direct Haplotype Resolution." American Journal of Human Genetics **46**(6): 1149-1155.
The International HapMap, C. (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.
Weir, B. S. (1996). Genetic Data Analysis II. Sunderland, MA., Sinauer Associates.
Wijsman, E. M. (1987). "A Deductive Method of Haplotype Analysis in Pedigrees." American Journal of Human Genetics **41**(3): 356-373.

Woolley, A. T., C. Guillemette, C. L. Cheung, D. E. Housman and C. M. Lieber (2000). "Direct haplotyping of kilobase-size DNA using carbon nanotube probes." Nature Biotechnology **18**(7): 760-763.

Zhang, H., H. C. Yang and Y. N. Yang (2008). "PoooL: an efficient method for estimating haplotype frequencies from large DNA pools." Bioinformatics **24**(17): 1942-1948.

Zhang, Q. F., Y. Z. Zhao, G. L. Chen and Y. Xu (2006). "Estimate haplotype frequencies in pedigrees." Bmc Bioinformatics **7**: 12.

Zhu, W. S., W. K. Fung and J. H. Guo (2007). "Incorporating genotyping uncertainty in haplotype frequency estimation in pedigree studies." Human Heredity **64**(3): 172-181.

Zhu, W. S., A. Y. C. Kuk and J. H. Guo (2009). "Haplotype Inference for Population Data with Genotyping Errors." Biometrical Journal **51**(4): 644-658.

Figure 1. Difference between true and estimated frequency of one of double heterozygous ($P_{AB|ab}$) according to $k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}$ for equal haplotype frequency setting $(0.25,\ 0.25,\ 0.25,\ 0.25)$

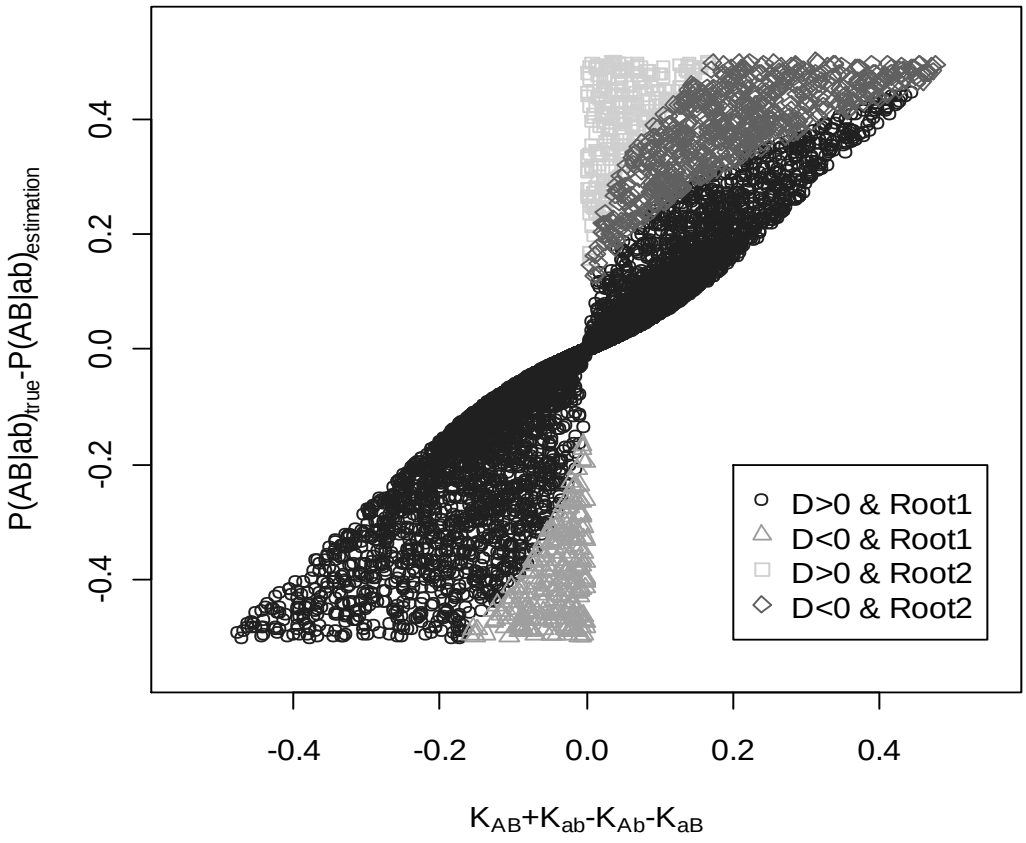


Figure A1. 2. Difference between true and estimated frequency of one of double heterozygous ($P_{AB|ab}$) according to $k_{\mathrm{AB}} + k_{\mathrm{ab}} - k_{\mathrm{Ab}} - k_{\mathrm{aB}}$ for unequal haplotype frequency setting (0.1, 0.2, 0.3, 0.4)

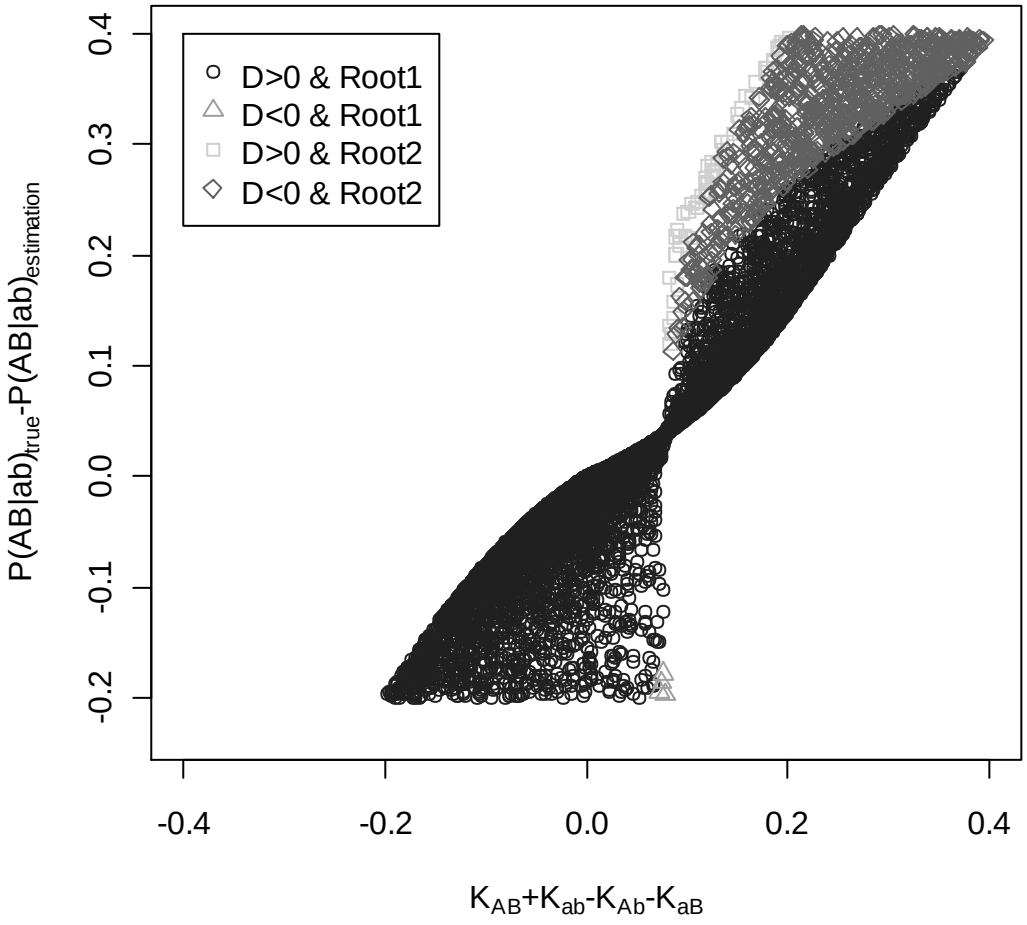


Figure A1. 3. EM estimated vs. true genotype frequency of one of double ($P_{AB|ab}$) for different levels of sum of double heterozygous genotypes for equal haplotype frequency setting (0.25, 0.25, 0.25, 0.25)

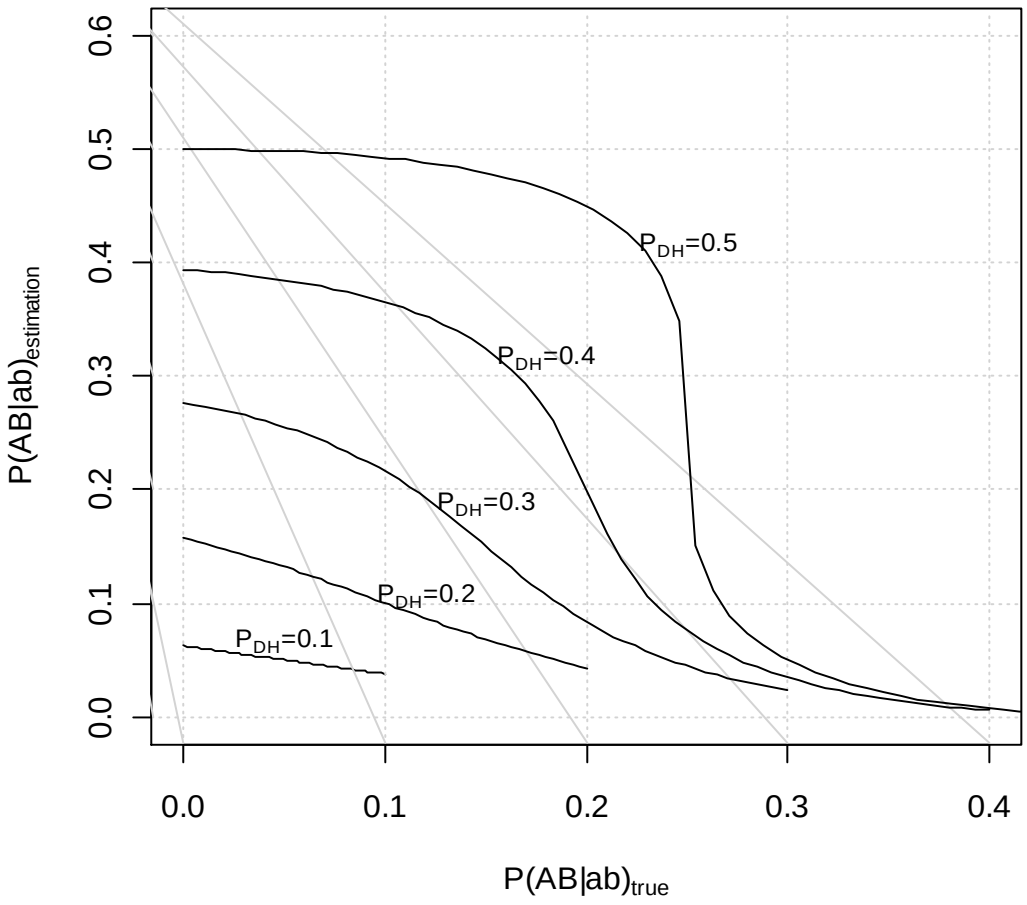


Figure A1. 4. EM estimated vs. true genotype frequency of one of double ($P_{AB|ab}$) for different levels of sum of double heterozygous genotypes for unequal haplotype frequency setting (0.1, 0.2, 0.3, 0.4)

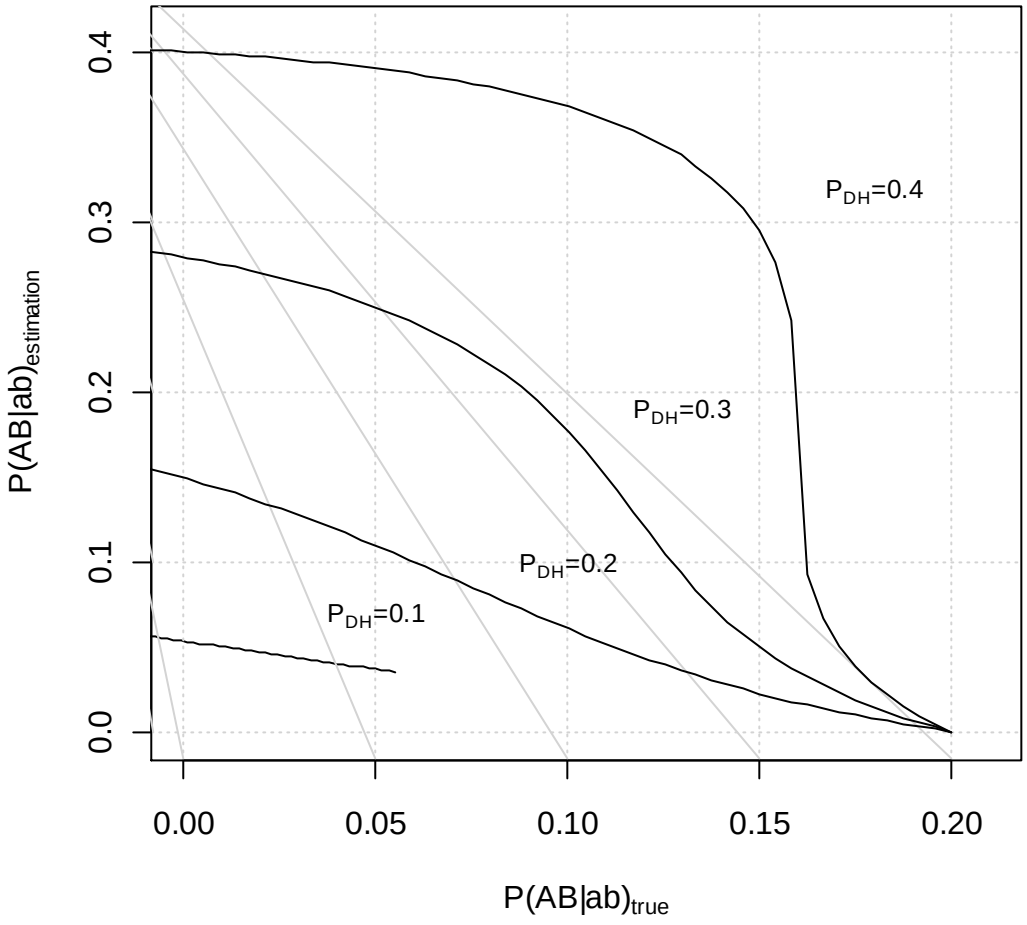


Figure A1. 5. Mean squared error of haplotype and genotype estimation by sum of double heterozygous genotype frequencies for equal haplotype frequency setting (100,000 population genotypes)

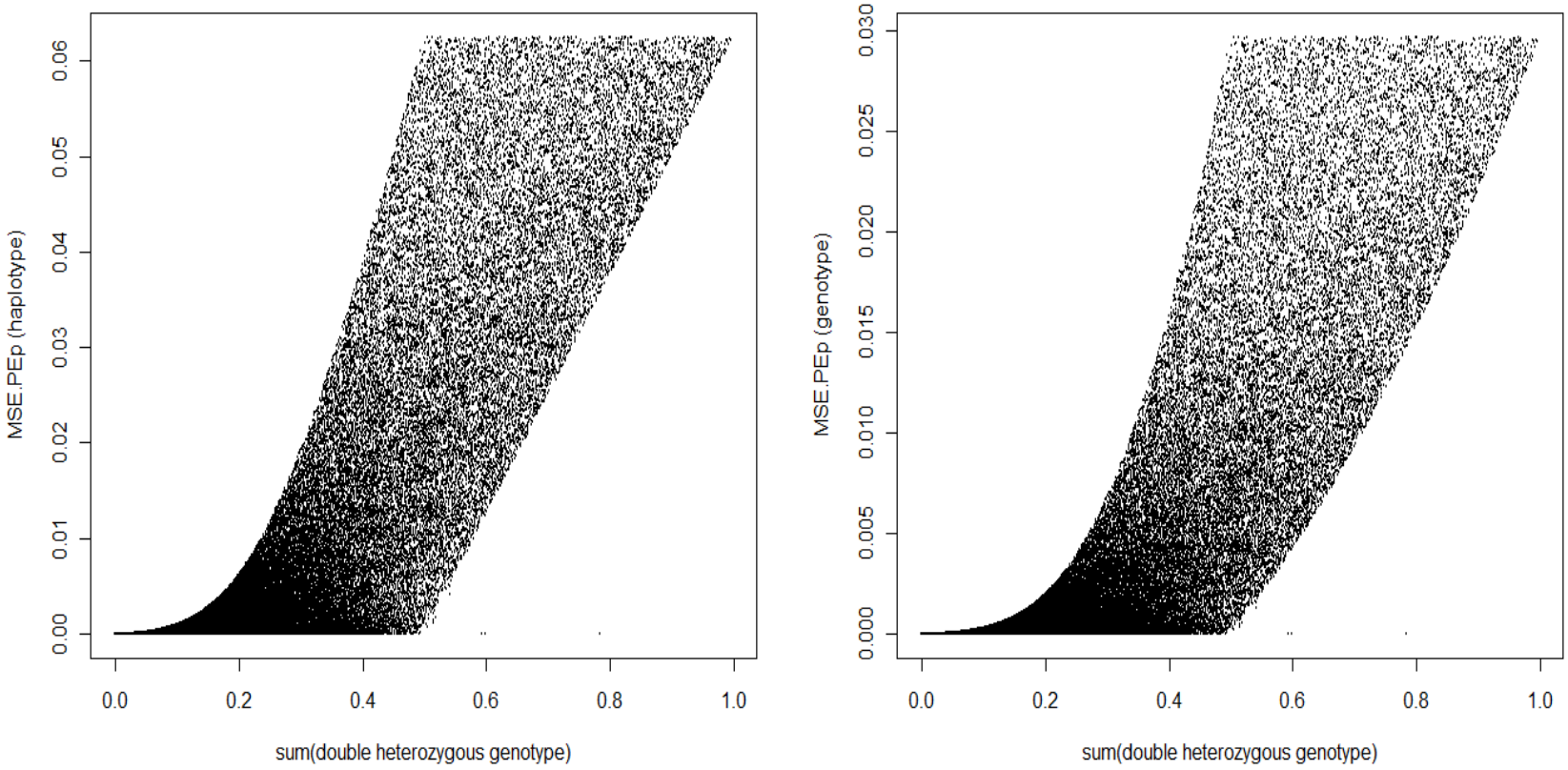
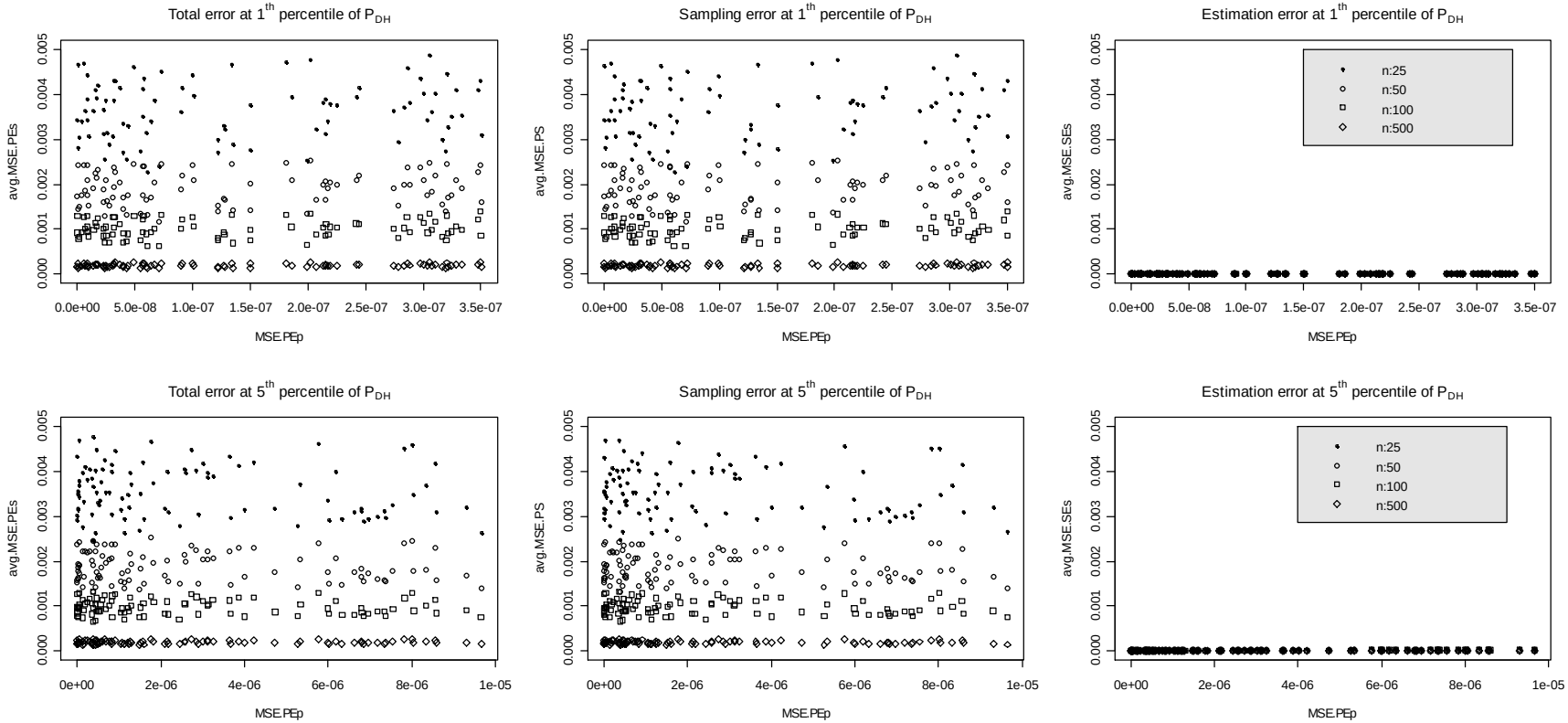


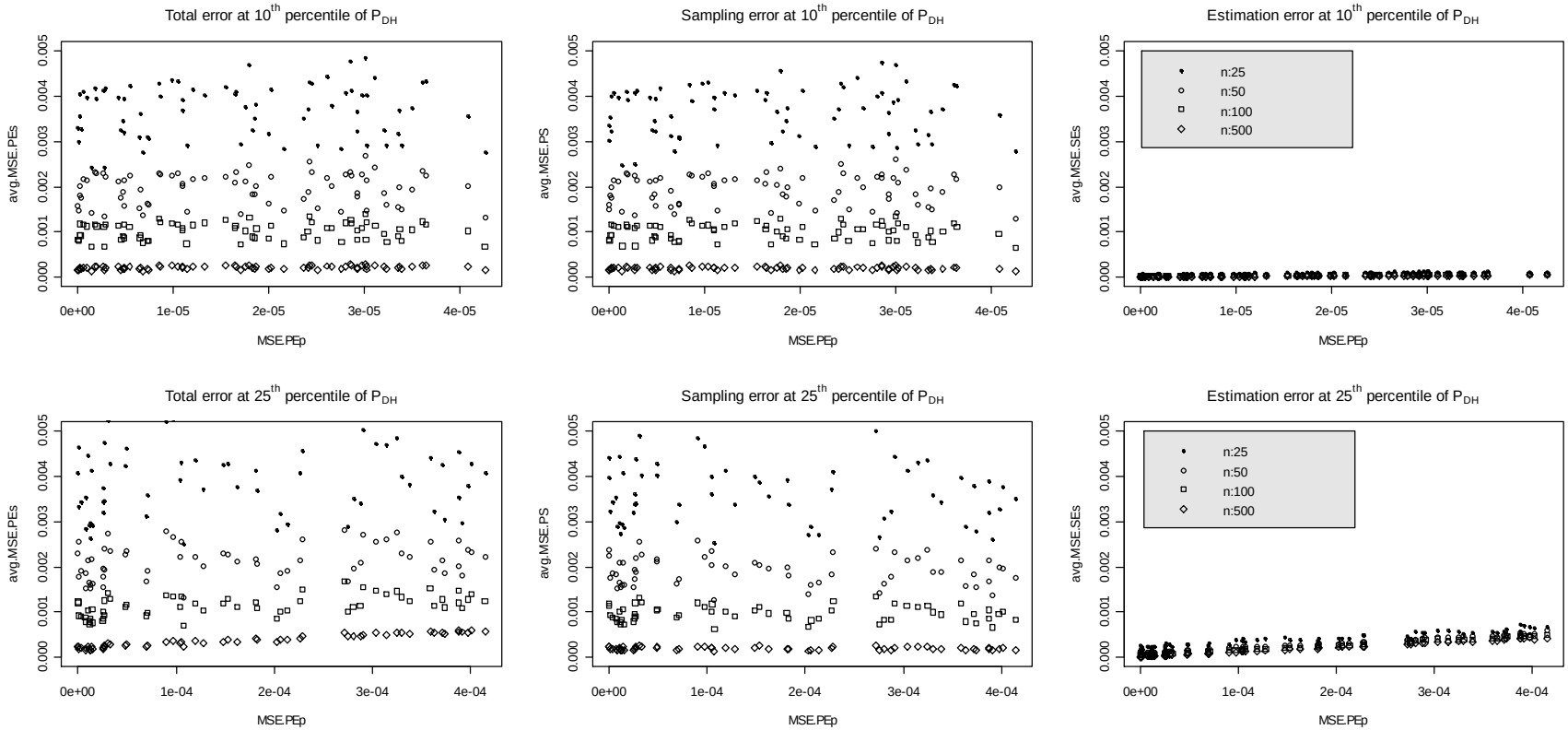
Figure A1. 6. Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for equal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 1st percentile of sum of double heterozygous genotypes: 0.002343

5th percentile of sum of double heterozygous genotypes: 0.012051

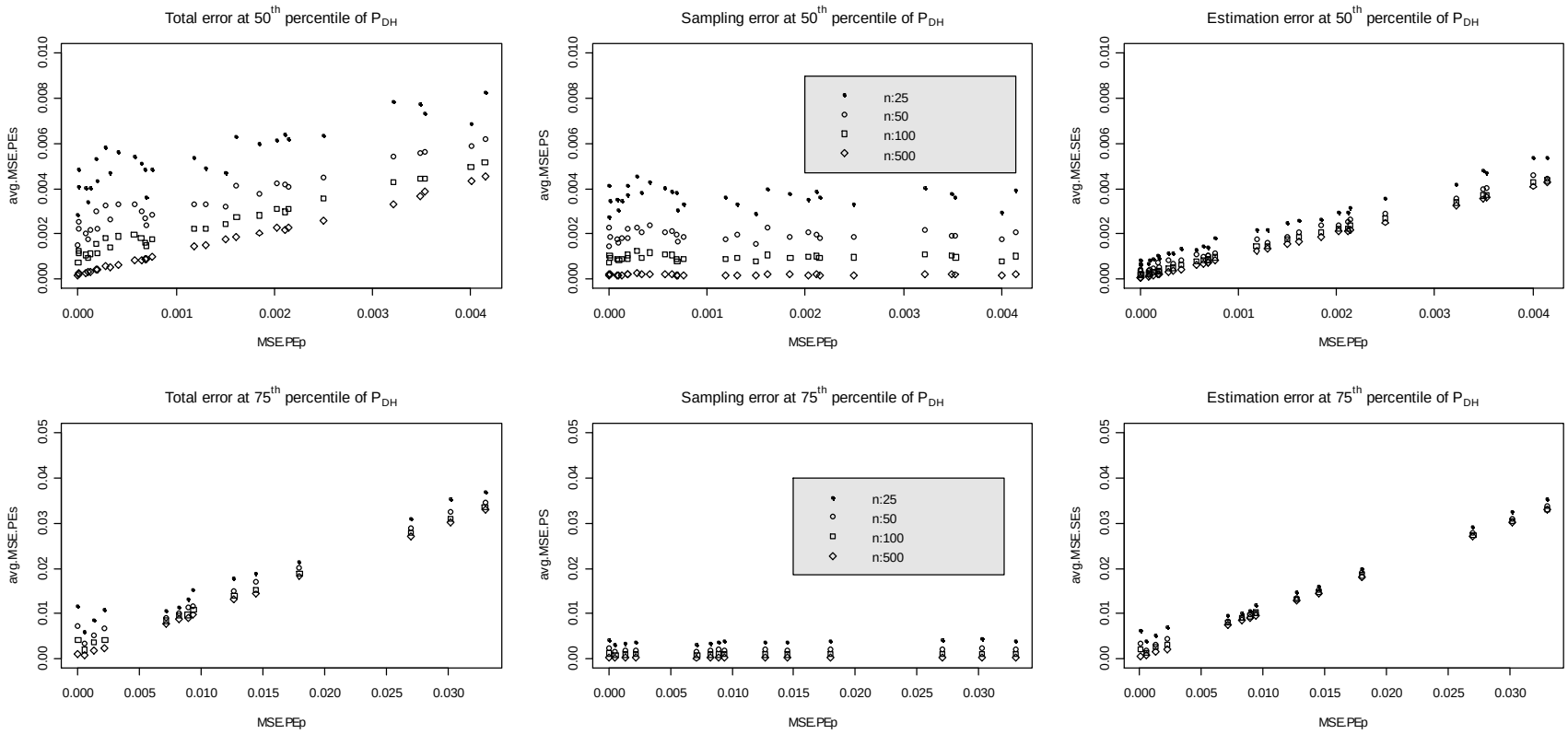
Figure A1. 6. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for equal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 10th percentile of sum of double heterozygous genotypes: 0.024944

25th percentile of sum of double heterozygous genotypes: 0.070267

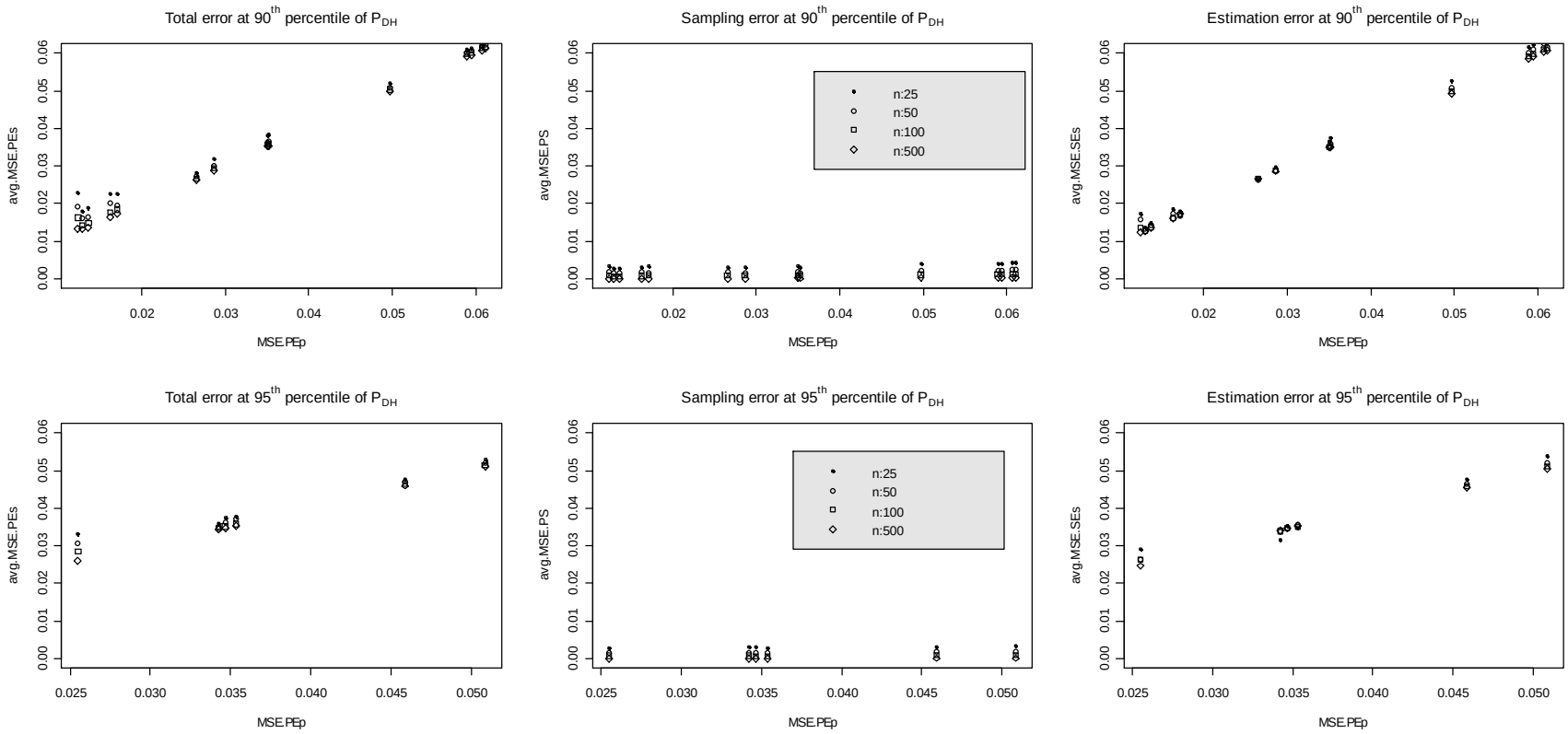
Figure A1. 6. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for equal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 50th percentile of sum of double heterozygous genotypes: 0.180454

75th percentile of sum of double heterozygous genotypes: 0.380338

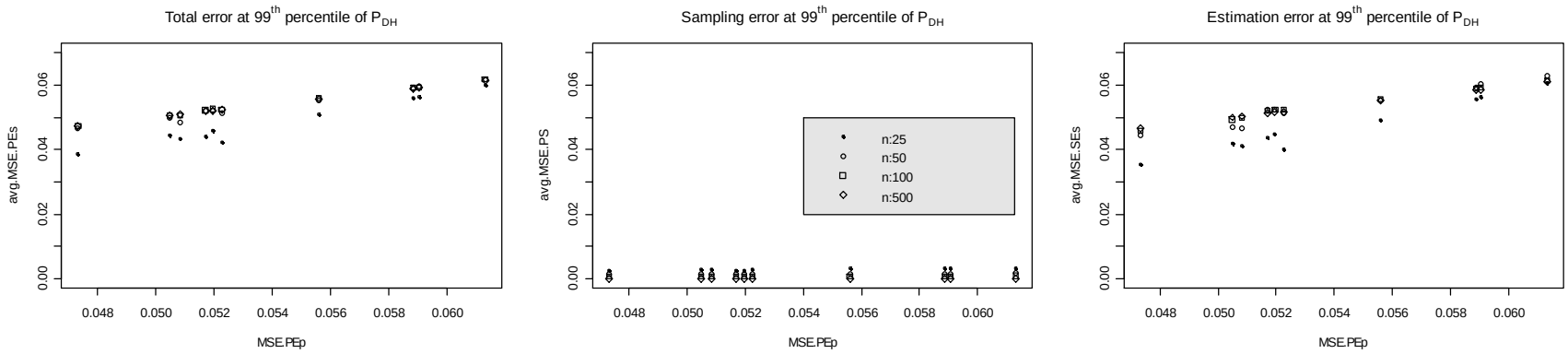
Figure A1. 6. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for equal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 90th percentile of sum of double heterozygous genotypes: 0.579567

95th percentile of sum of double heterozygous genotypes: 0.692275

Figure A1. 6. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for equal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 99th percentile of sum of double heterozygous genotypes: 0.857524

Figure A1. 7. Averages ± Standard deviations of mean squared errors against MSE_{PEp} at sample size 100 for equal haplotype frequency setting

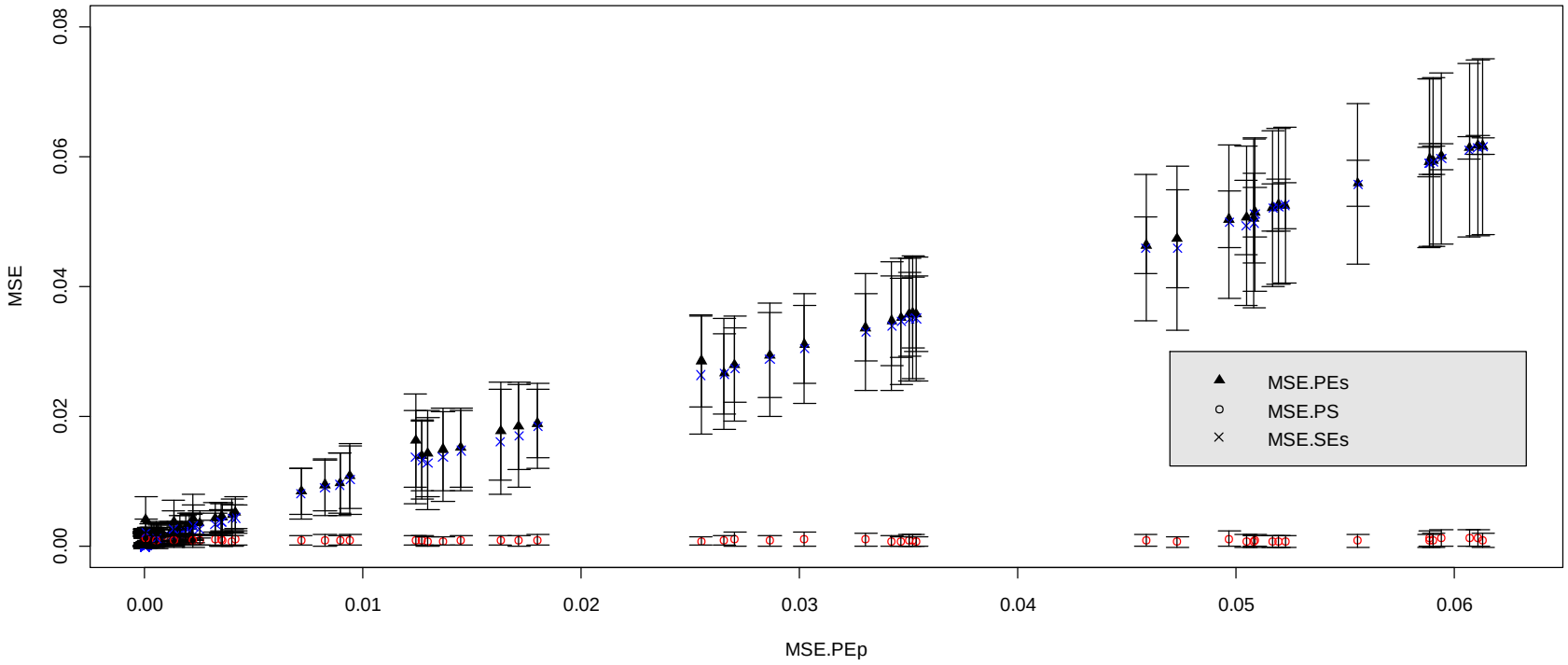


Figure A1. 8. Mean squared error of estimation for haplotype or genotype against heteozygosity for equal frequency setting (100,000 population genotypes)

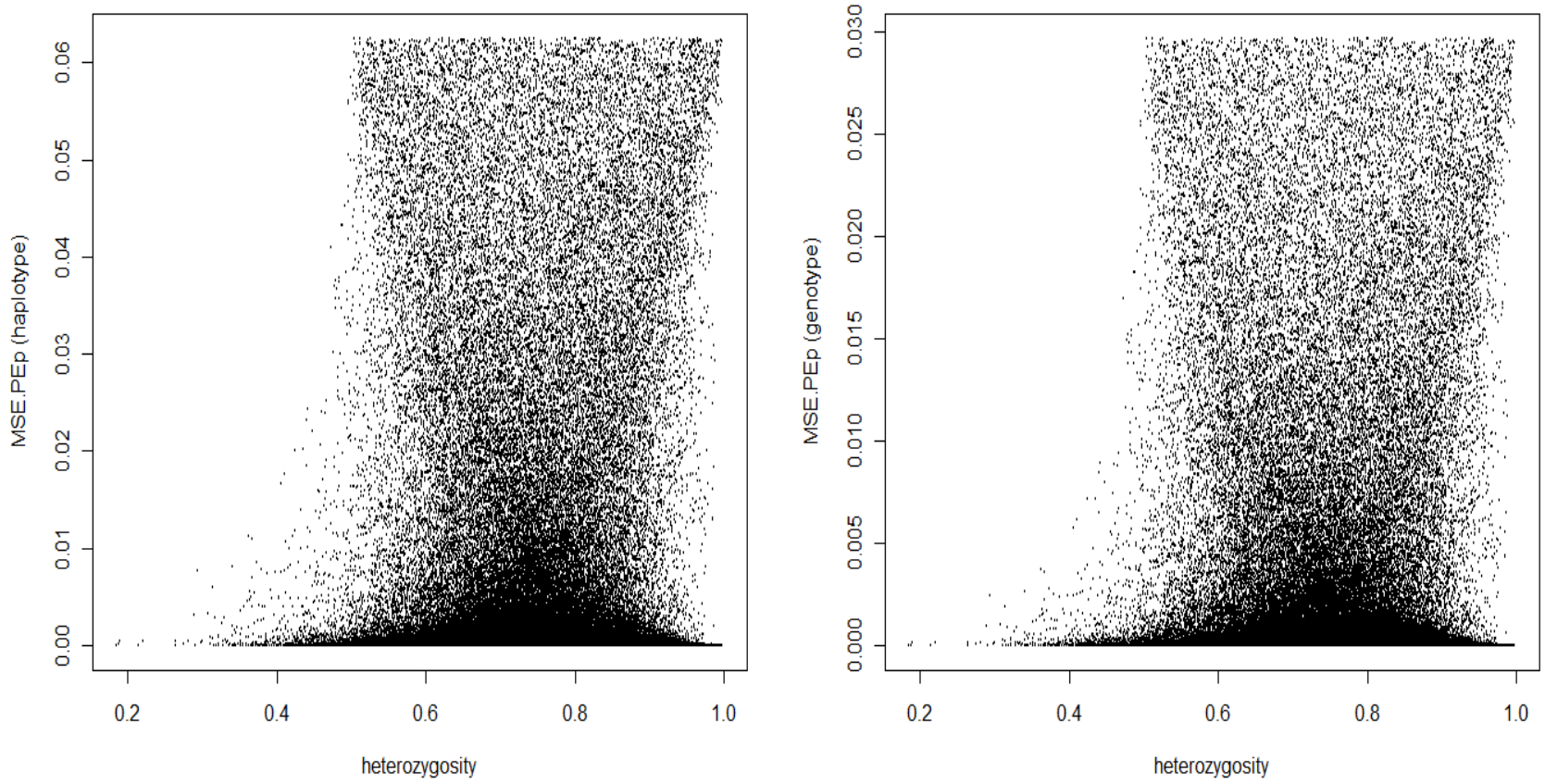


Figure A1. 9. Mean squared error of haplotype and genotype estimation for by sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (100,000 population genotypes)

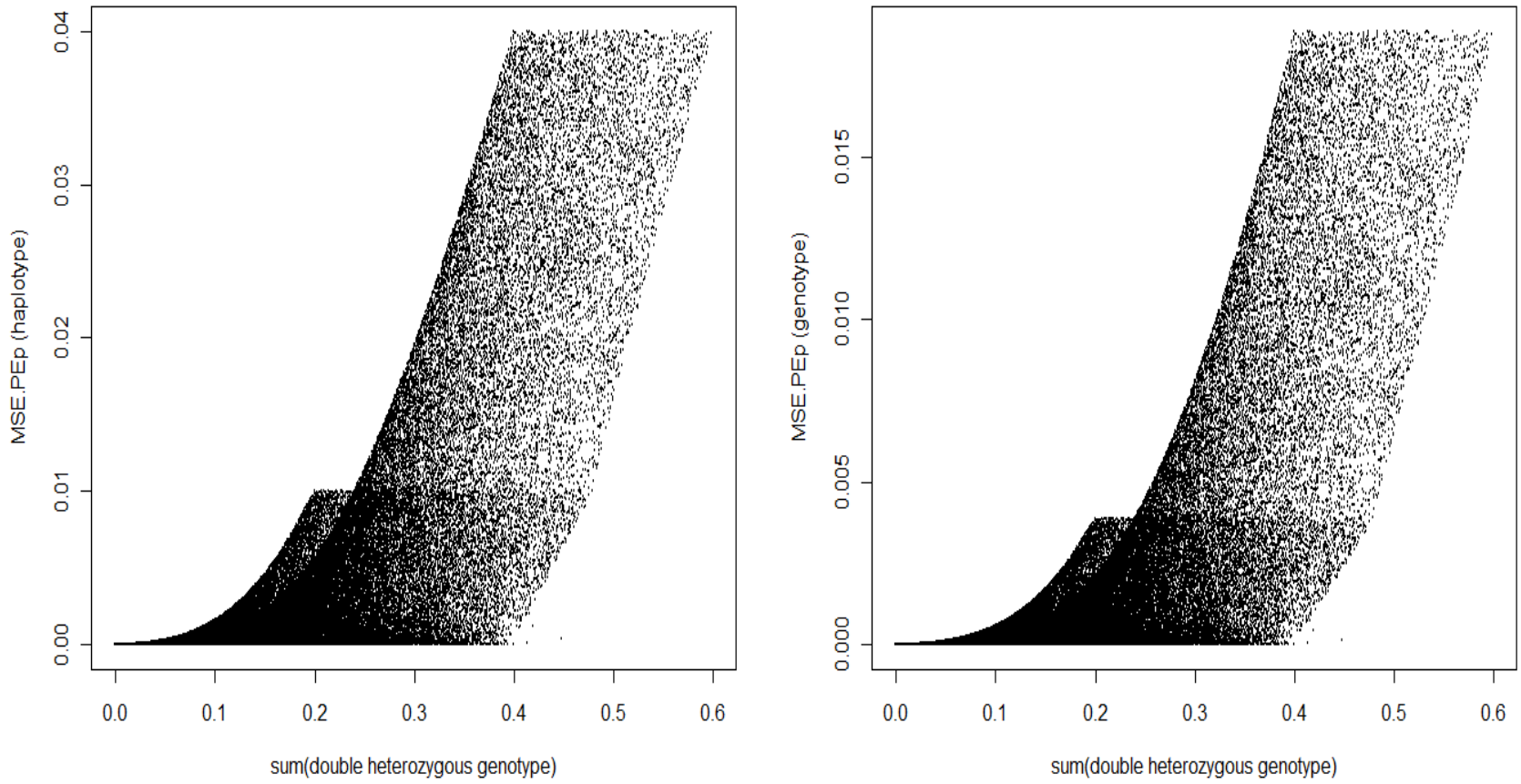
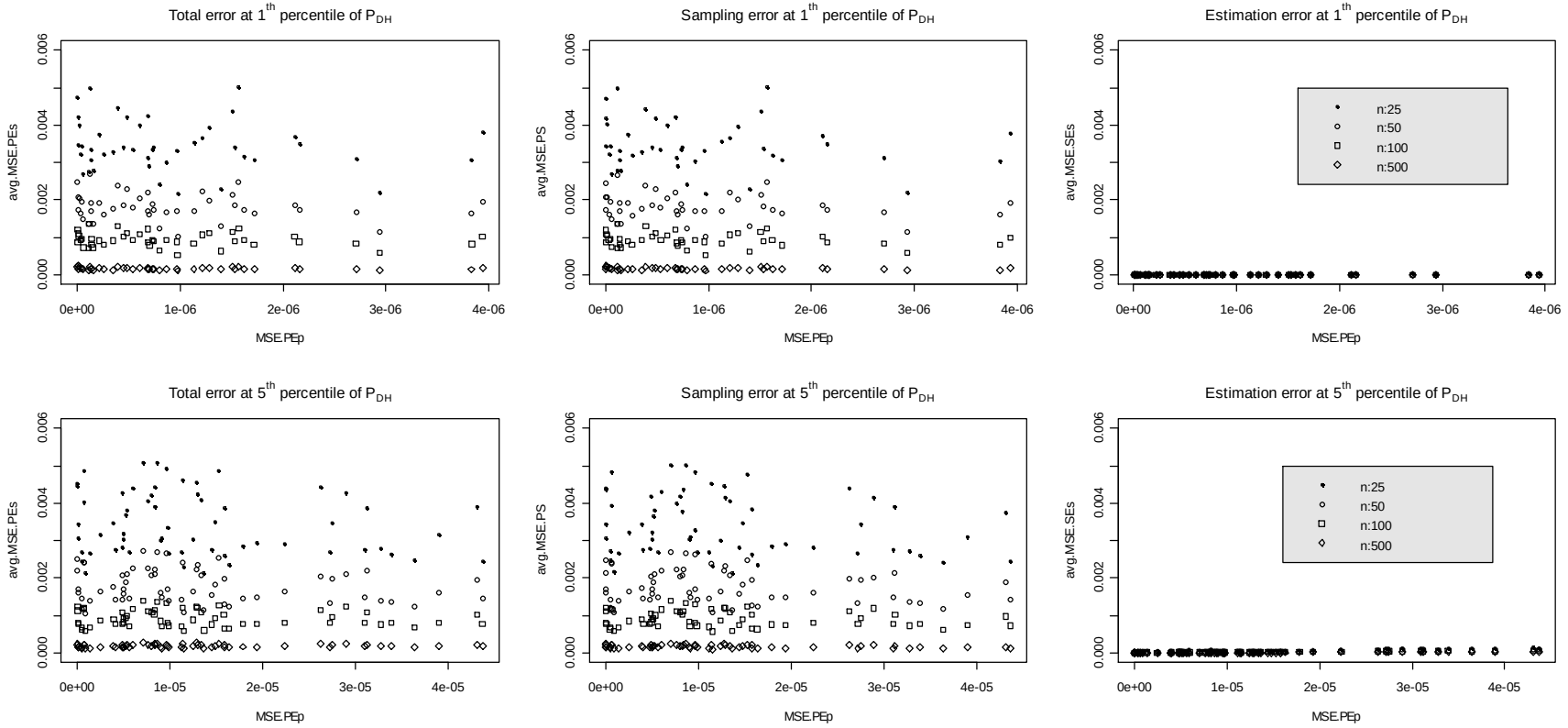


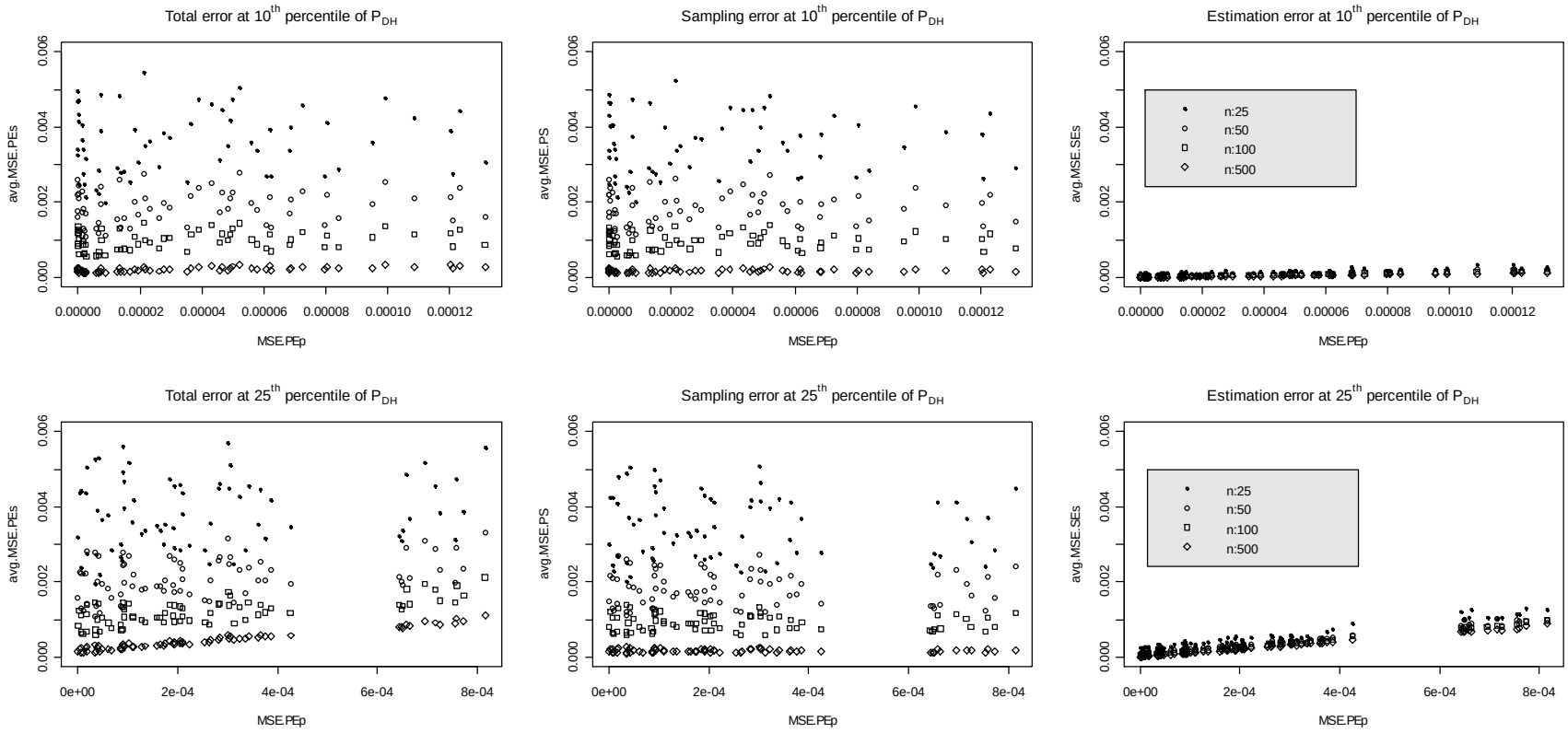
Figure A1. 10. Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 1st percentile of sum of double heterozygous genotypes: 0.006467

5th percentile of sum of double heterozygous genotypes: 0.020859

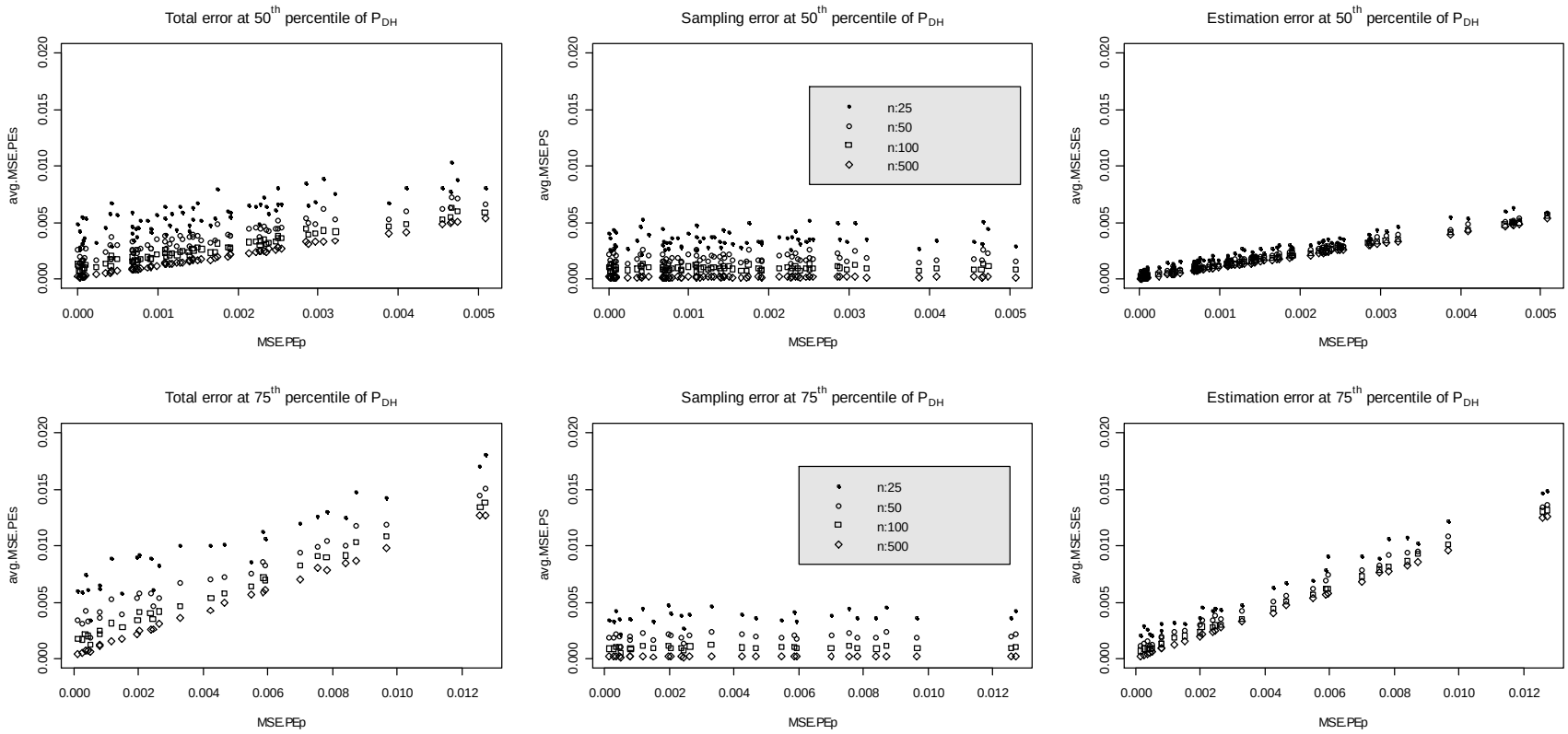
Figure A1. 10. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 10th percentile of sum of double heterozygous genotypes: 0.035951

25th percentile of sum of double heterozygous genotypes: 0.078705

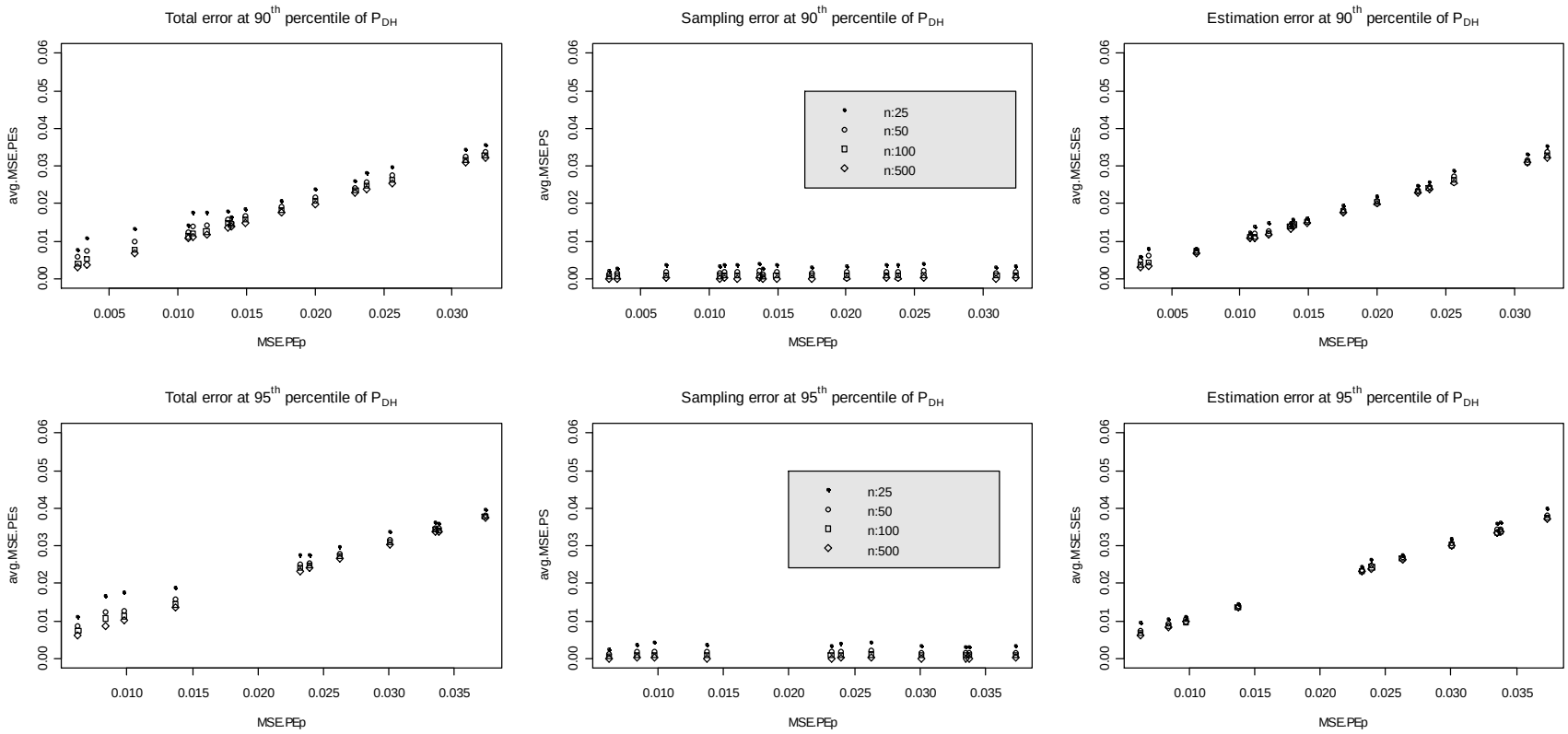
Figure A1. 10. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 50th percentile of sum of double heterozygous genotypes: 0.156601

75th percentile of sum of double heterozygous genotypes: 0.259721

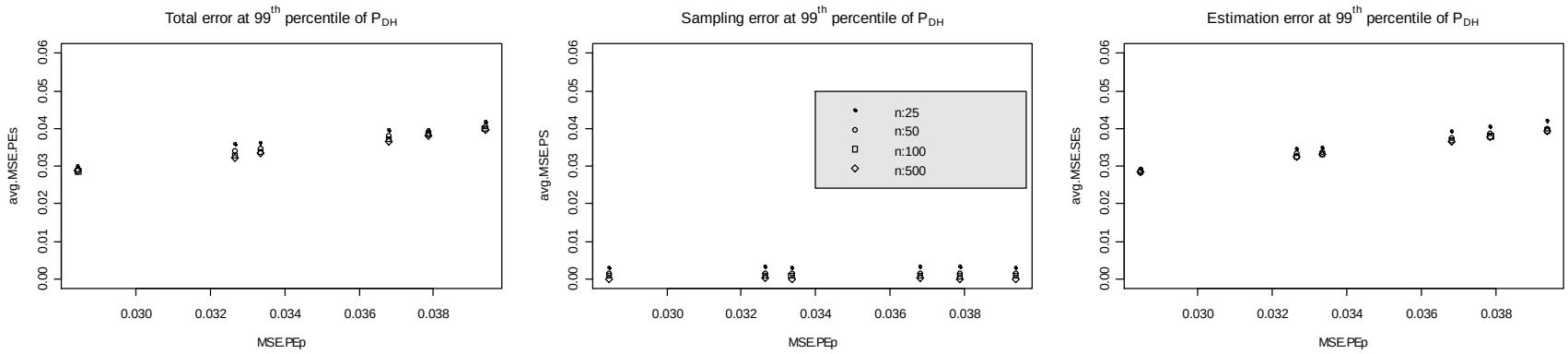
Figure A1. 10. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 90th percentile of sum of double heterozygous genotypes: 0.366842

95th percentile of sum of double heterozygous genotypes: 0.421101

Figure A1. 10. (continued) Average mean squared errors at different percentiles of sum of double heterozygous genotype frequencies for unequal haplotype frequency setting (sample size: 25, 50, 100, 200, replicates: 200 each)



Note: 99th percentile of sum of double heterozygous genotypes: 0.509127

Figure A1. 11. Averages ± Standard deviations of mean squared errors against MSE_{PEp} at sample size 100 for unequal haplotype frequency setting

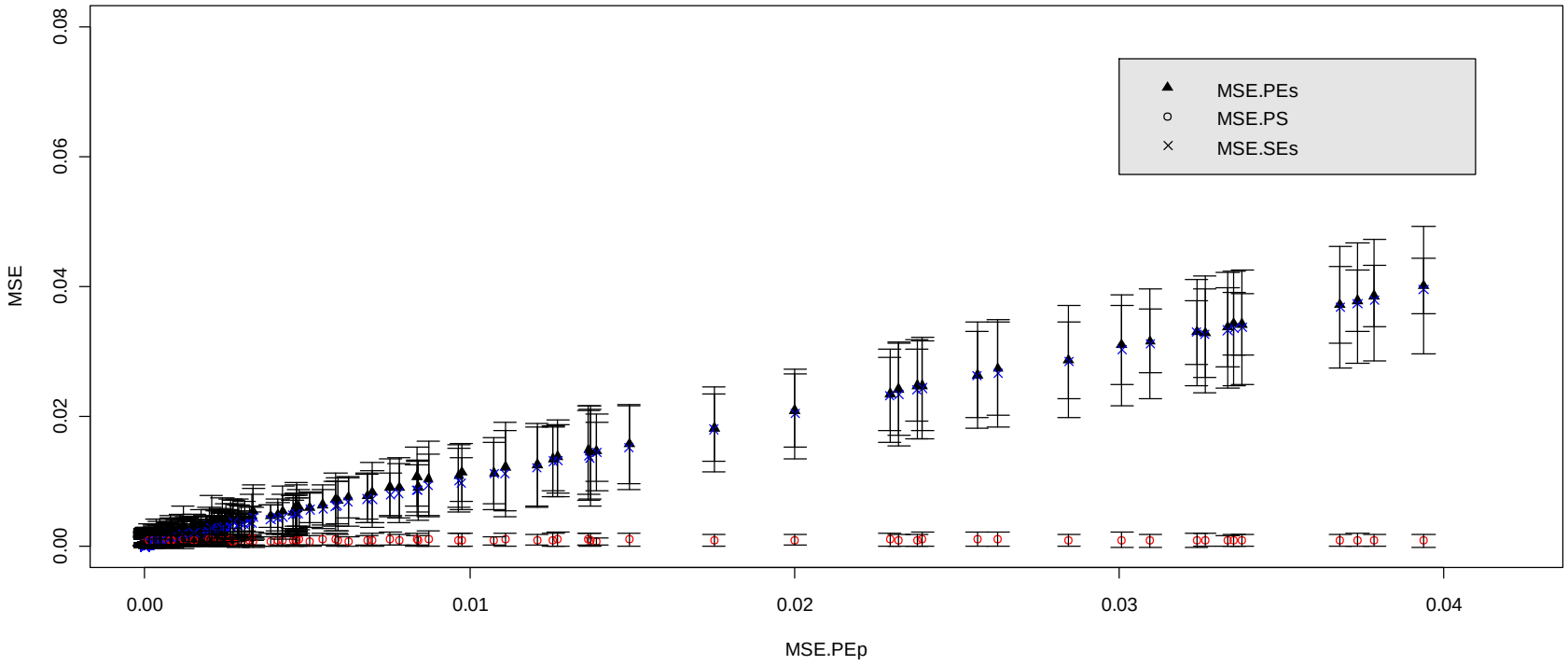


Figure A1. 12. Mean squared error of estimation for haplotype or genotype against heteozygosity for unequal frequency setting (100,000 population genotypes)

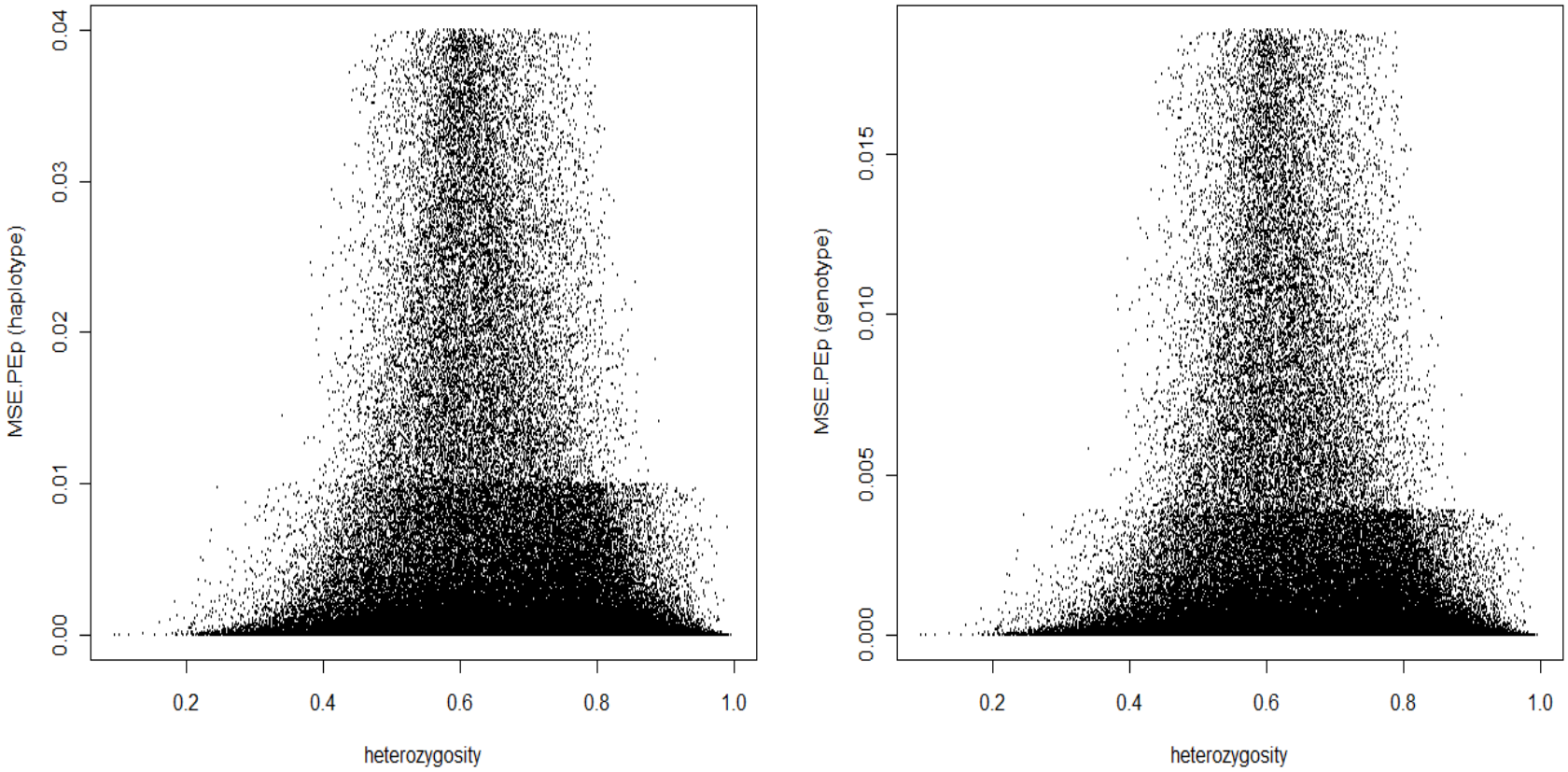


Figure A1. 13. Averages ± Standard deviations of MSE.PEp from HWD-ECM against MSE.PEp from EM algorithm with 5000 initial haplotypes for 50 randomly selected genotype sets of equal haplotype frequency setting

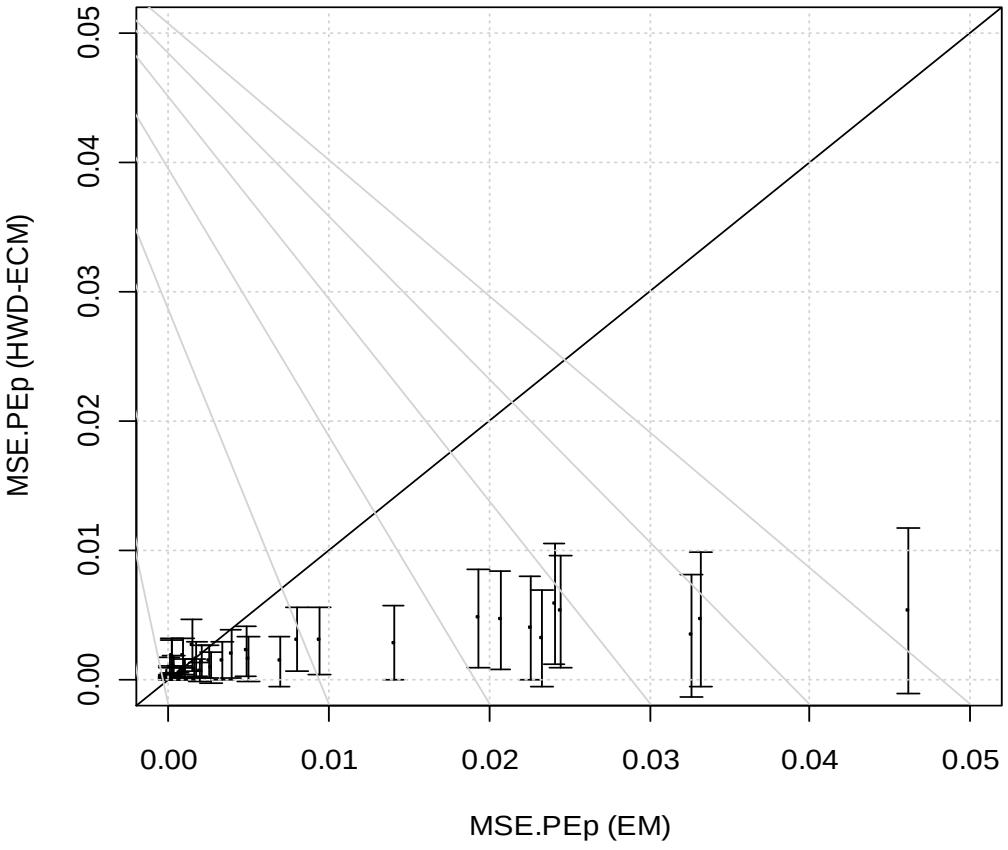


Figure A1. 14. Scatter plot of average HWD-ECM MSE.PEp s sorted according to EM MSE.PEp from 25, 50, 100 and 200 initial haplotypes for 50 randomly selected genotype sets with equal population haplotype frequency setting

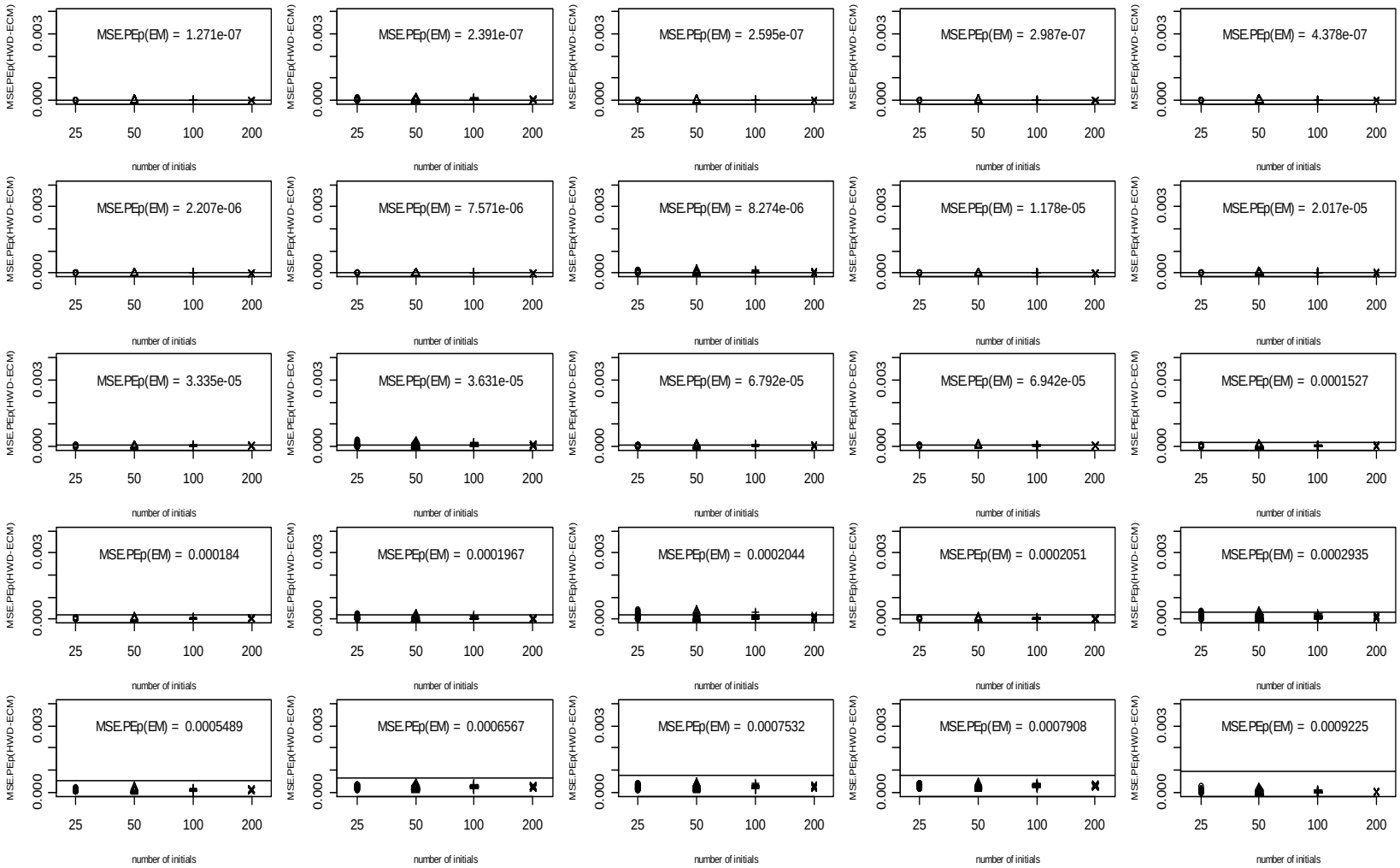


Figure A1. 14. (Continued) Scatter plot of average HWD-ECM MSE.PEp s sorted according to EM MSE.PEp from 25, 50, 100 and 200 initial haplotypes for 50 randomly selected genotype sets with equal population haplotype frequency setting

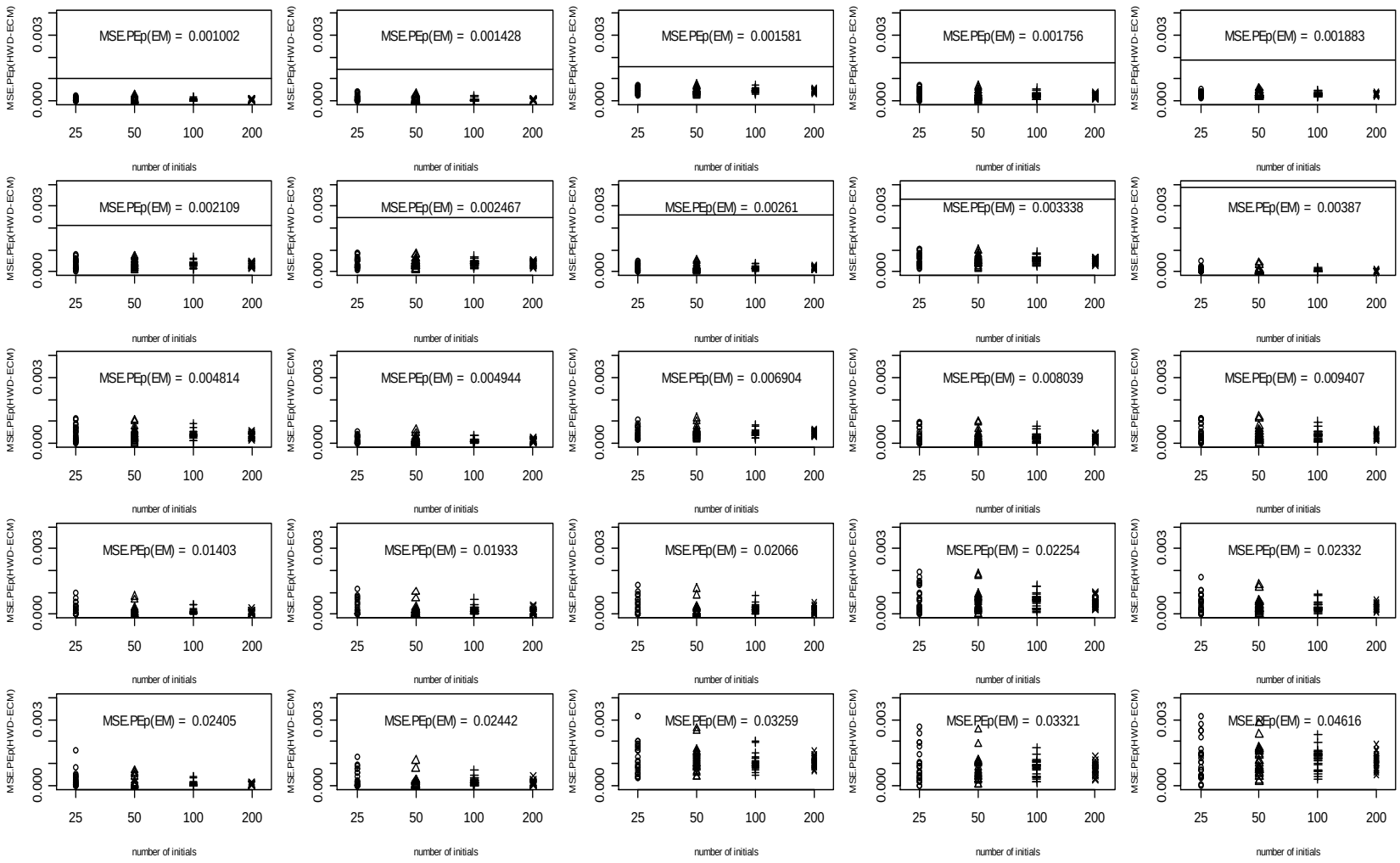


Figure A1. 15. Average ± Standard deviation of MSE.PEp from HWD-ECM against MSE.PEp from EM algorithm with 5000 initial haplotypes for 50 randomly selected genotype sets of unequal haplotype frequency setting (0.1, 0.2, 0.3, 0.4)

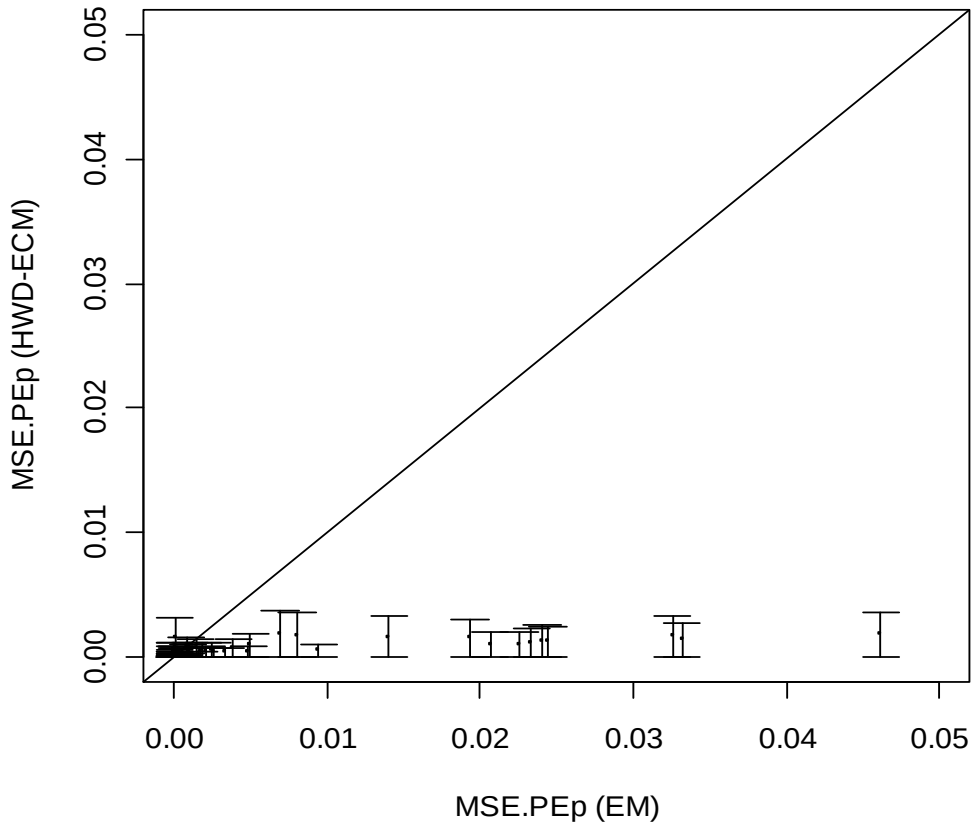


Figure A1. 16. Scatter plot of average HWD-ECM MSE.PEp s sorted according to EM MSE.PEp from 25, 50, 100 and 200 initial haplotypes for 50 randomly selected genotype sets with unequal population haplotype frequency setting

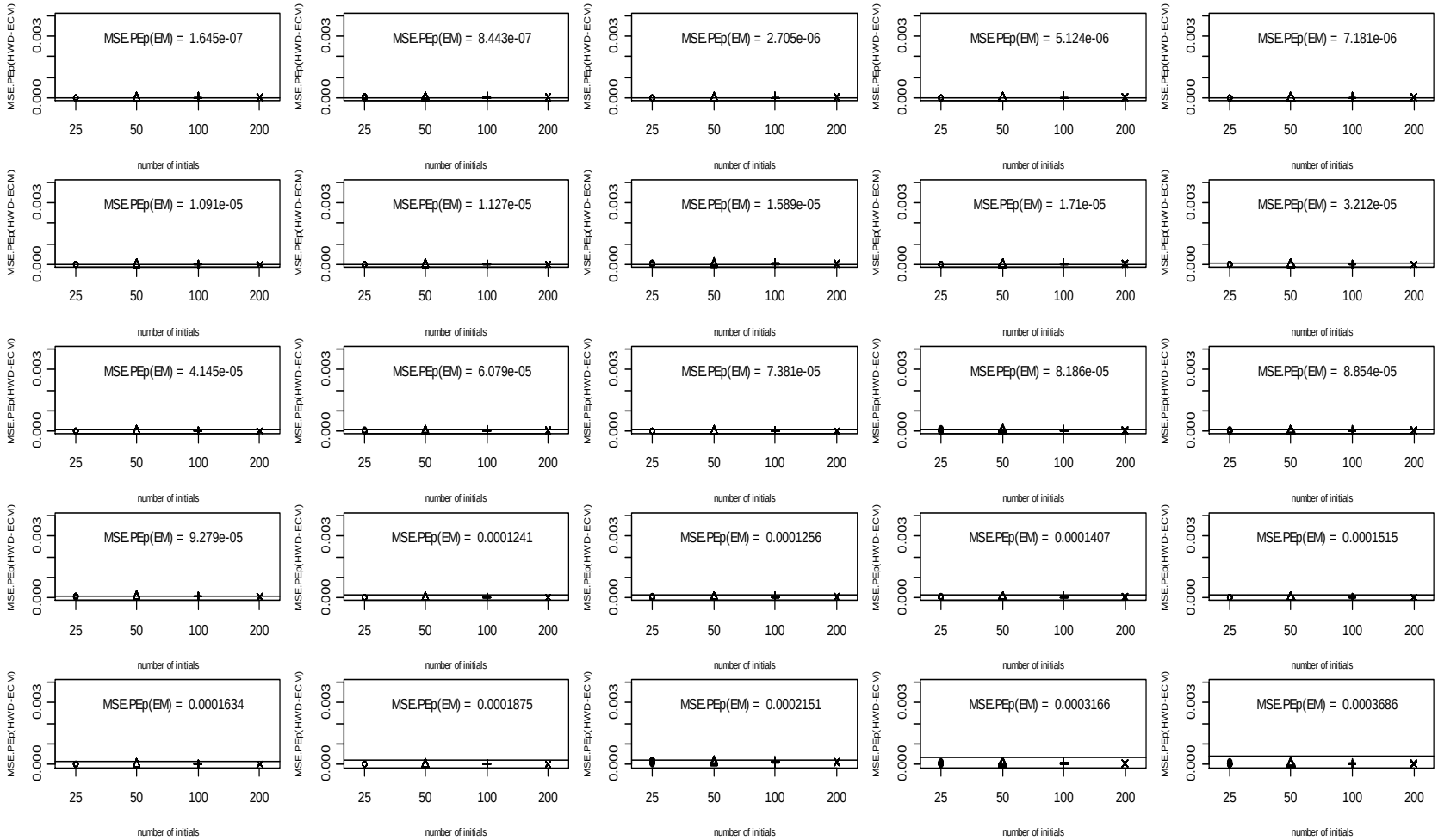


Figure A1. 16. (Continued) Scatter plot of average HWD-ECM MSE.PEp s sorted according to EM MSE.PEp from 25, 50, 100 and 200 initial haplotypes for 50 randomly selected genotype sets with unequal population haplotype frequency setting

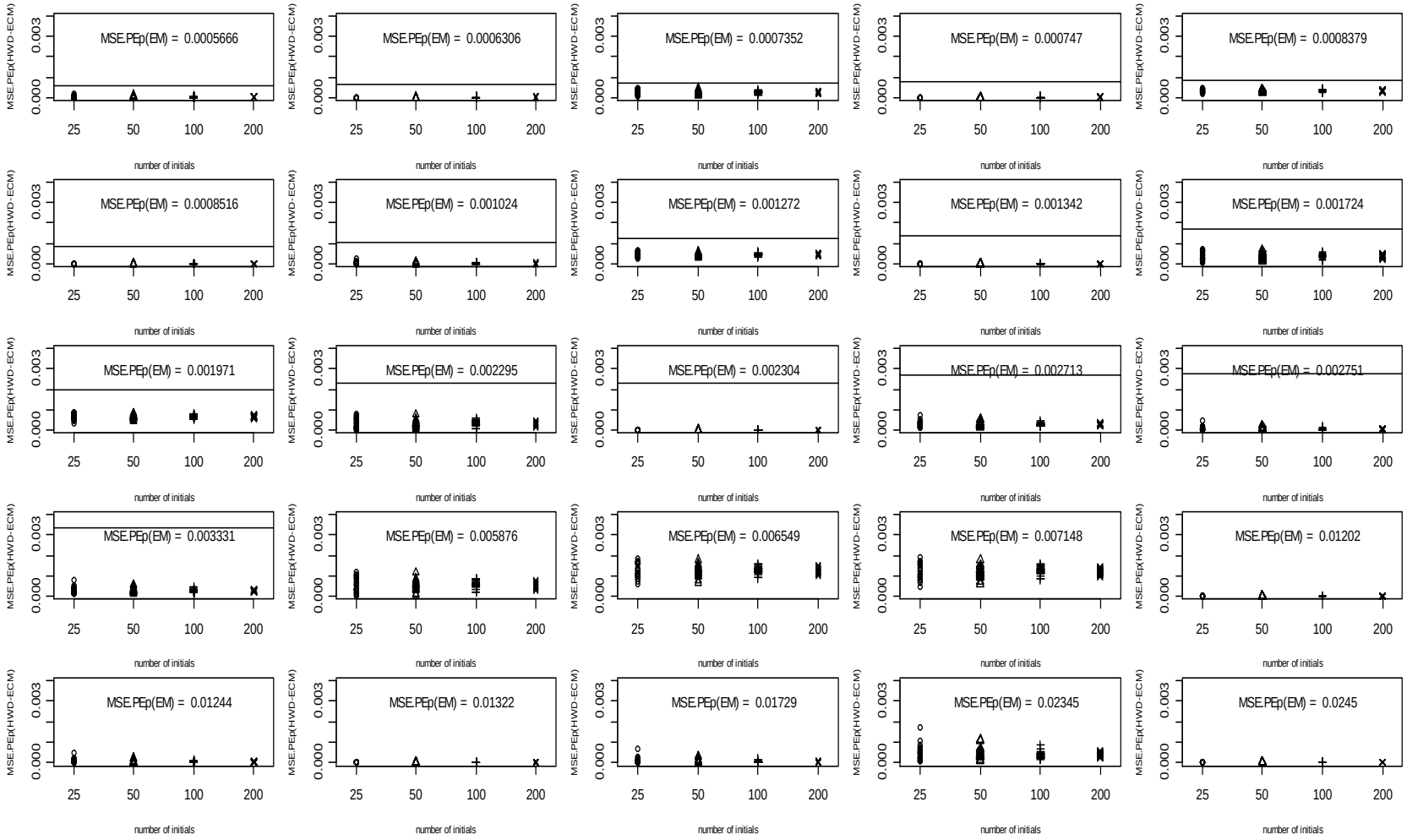
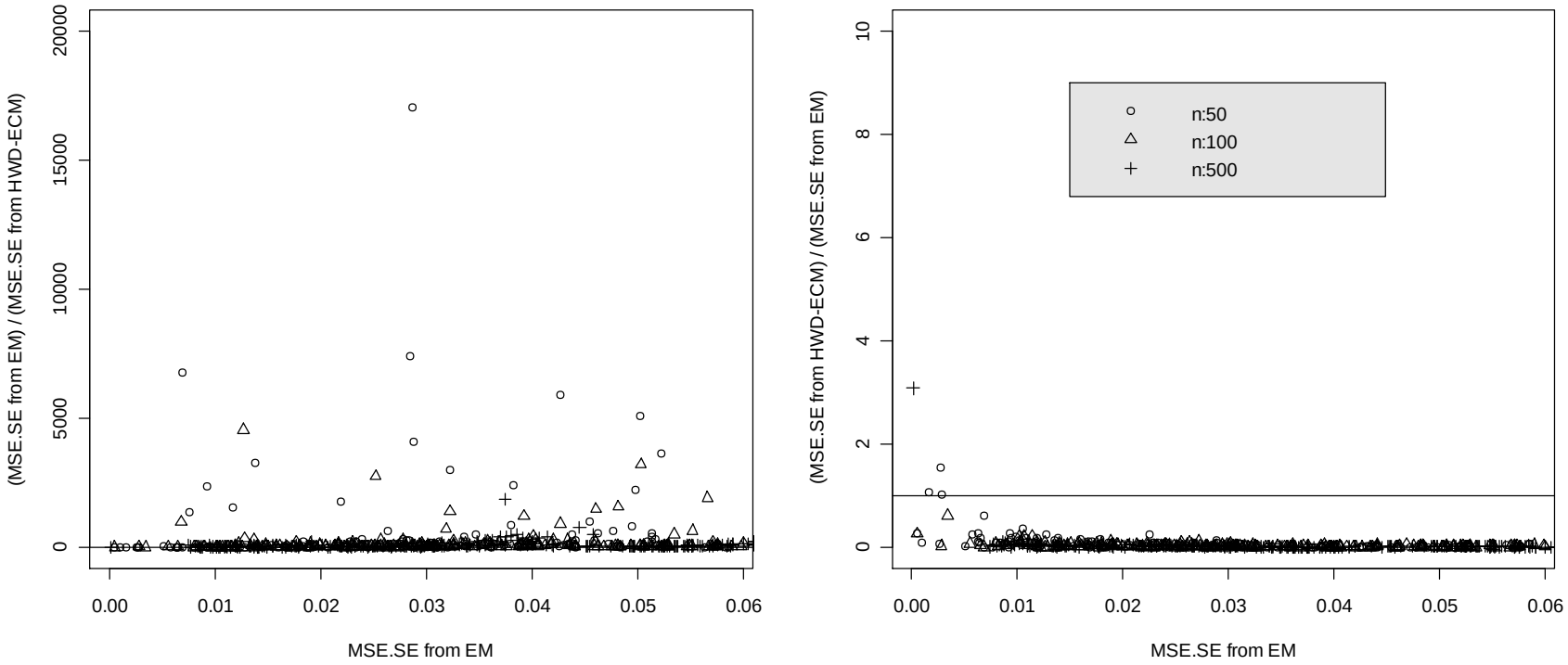
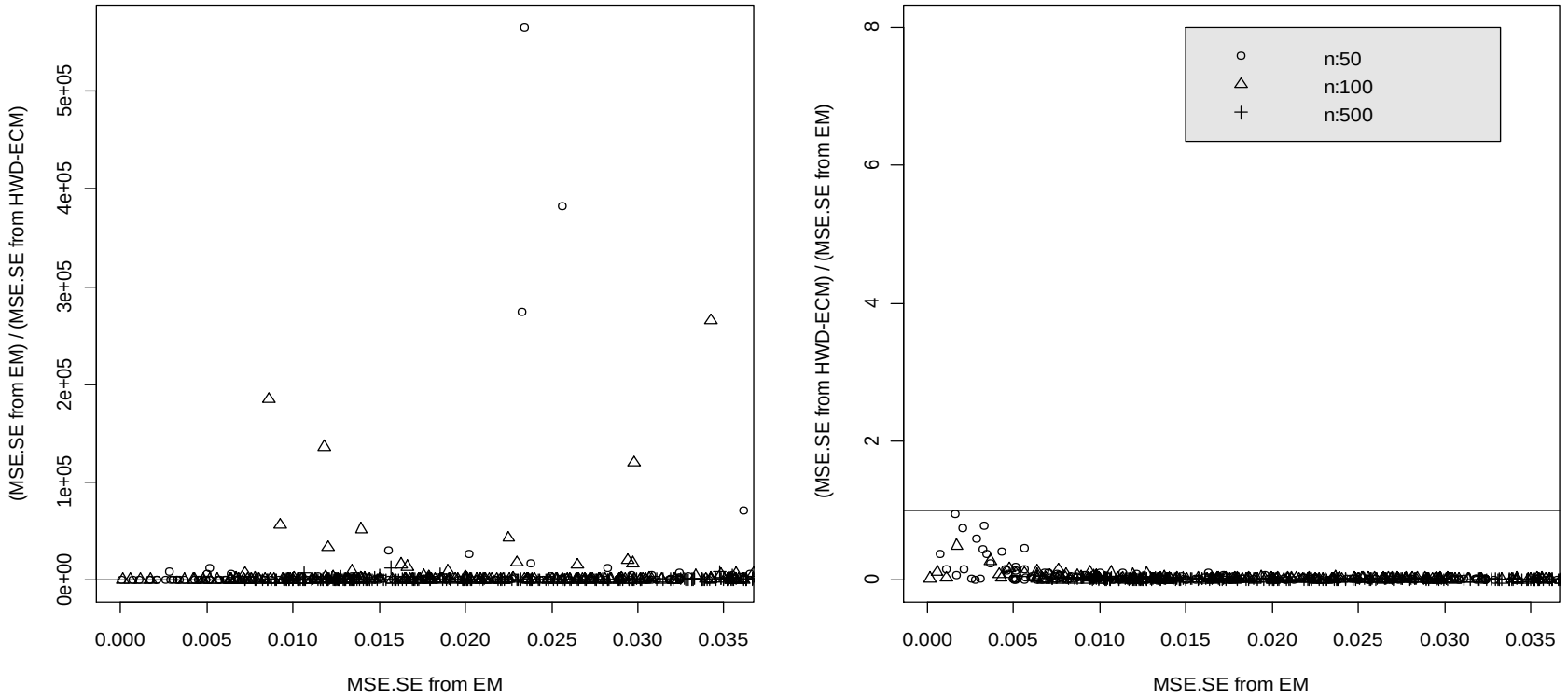


Figure A1. 17. Comparison of HWD-ECM MSE.SE from against EM MSE.SE for five genotype settings with different levels of MSE.PEp for equal haplotype frequency setting (left: (EM MSE.SE)/ (HWD-ECM MSE.SE) vs. MSE.SE from EM, right: (HWD-ECM MSE.SE) / (HWD-ECM MSE.SE)



Note: sample size: 50, 100, 500 (50 replicates each)

Figure A1. 18. Comparison of HWD-ECM MSE.SE from against EM MSE.SE for five genotype settings with different levels of MSE.PEp for equal haplotype frequency setting (left: (EM MSE.SE)/ (HWD-ECM MSE.SE) vs. MSE.SE from EM, right: (HWD-ECM MSE.SE) / (HWD-ECM MSE.SE)



Note: sample size: 50, 100, 500 (50 replicates each)

Figure A1. 19. All possible genotype features of 3 SNPs

	"ABC"	"aBC"	"AbC"	"abC"	"ABc"	"aBc"	"Abc"	"abc"
"ABC"	H11	S12	S13	D14	S15	D16	D17	T18
"aBC"		H22	D23	S24	D25	S26	T27	D28
"AbC"			H33	S34	D35	T36	S37	D38
"abC"				H44	T45	D46	D47	S48
"ABc"					H55	S56	S57	D58
"aBc"						H66	D67	S68
"Abc"							H77	S78
"abc"								H88

H: homozygous genotype
S: single heterozygous genotype
D: double heterozygous genotype
T: triple heterozygous genotype

Figure A1. 20. Genotype features by fixing particular allele for each SNP

1. Fixing “A” for first SNP,

	"ABC"	"AbC"	"ABc"	"Abc"
"ABC"	H11	S13	S15	D17
"AbC"		H33	D35	S37
"ABc"			H55	S57
"Abc"				H77

2. Fixing “a” for first SNP,

	"aBC"	"abC"	"aBc"	"abc"
"aBC"	H22	S24	S26	D28
"abC"		H44	D46	S48
"aBc"			H66	S68
"abc"				H88

3. Fixing “B” for second SNP,

	"ABC"	"aBC"	"ABc"	"aBc"
"ABC"	H11	S12	S15	D16
"aBC"		H22	D25	S26
"ABc"			H55	S56
"aBc"				H66

4. Fixing “b” for second SNP,

	"AbC"	"abC"	"Abc"	"abc"
"AbC"	H33	S34	S37	D38
"abC"		H44	D47	S48
"Abc"			H77	S78
"abc"				H88

5. Fixing “C” for third SNP,

	"ABC"	"aBC"	"AbC"	"abC"
"ABC"	H11	S12	S13	D14
"aBC"		H22	D23	S24
"AbC"			H33	S34
"abC"				H44

6. Fixing “c” for third SNP,

	"ABc"	"aBc"	"Abc"	"abc"
"ABc"	H55	S56	S57	D58
"aBc"		H66	D67	S68
"Abc"			H77	S78
"abc"				H88

Figure A1. 21. Genotype features by merging genotypes according to each SNP

1. Merging "A" or "a" for first SNP,

	"BC"	"bC"	"Bc"	"bc"
"BC"	H11+S12+H22	S13+D14+D23+S24	S15+D16+D25+S26	D17+T18+ T27+D28
"bC"		H33+S34+H44	D35+T36+ T45+D46	S37+D38+D47+S48
"Bc"			H55+S56+H66	S57+D58+D67+S68
"bc"				H77+S78+H88

2. Merging "B" or "b" for second SNP,

	"AC"	"aC"	"Ac"	"ac"
"AC"	H11+S13+H33	S12+D14+D23+S34	S15+D17+D35+S37	D16+T18+ T36+D38
"aC"		H22+S24+H44	D25+T27+ T45+D47	S26+D28+D46+S48
"Ac"			H55+S57+H77	S56+D58+D67+S78
"ac"				H66+S68+H88

3. Merging "C" or "c" for third SNP,

	"AB"	"aB"	"Ab"	"ab"
"AB"	H11+S15+H55	S12+D16+D25+S56	S13+D17+D35+S57	D14+T18+T45+D58
"aB"		H22+S26+H66	D23+T27+T36+D67	S24+D28+D46+S68
"Ab"			H33+S37+H77	S34+D38+D47+S78
"ab"				H44+S48+H88

Table 1. Summary of population and sampling settings based on nine percentiles of sum of double heterozygous genotype frequencies for equal (0.25, 0.25, 0.25 and 0.25) and unequal (0.1, 0.2, 0.3 and 0.4) haplotype setting

Population	percentile	1st	5th	10th	25th	50th	75th	90th	95th	99th
Equal haplotype frequency (0.25,0.25, 0.25,0.25): 100,000 HW deviation sets	sum of double heterozygous genotype ±0.0001	0.002343 ±0.0001	0.012051 ±0.0001	0.024944 ±0.0001	0.070267 ±0.0001	0.180454 ±0.0001	0.380338 ±0.0001	0.579567 ±0.0001	0.692275 ±0.0001	0.857524 ±0.0001
	n	87	96	73	60	30	14	14	6	10
	mean (MSE.PEp)	1.38E-07	2.75E-06	1.70E-05	0.000167	0.001324	0.01238	0.034835	0.037752	0.053939
	sd (MSE.PEp)	1.19E-07	2.86E-06	1.22E-05	0.000142	0.001302	0.011022	0.019527	0.009125	0.004533
Unequal haplotype frequency (0.1,0.2, 0.3, 0.4) : 100,000 HW deviation sets	sum of double heterozygous genotype ±0.0001	0.006497 ±0.0001	0.020859 ±0.0001	0.035951 ±0.0001	0.078705 ±0.0001	0.156601 ±0.0001	0.259721 ±0.0001	0.366842 ±0.0001	0.421101 ±0.0001	0.509127 ±0.0001
	n	44	61	62	71	70	28	16	11	6
	mean (MSE.PEp)	9.53E-07	1.32E-05	3.73E-05	0.000256	0.001626	0.004339	0.016414	0.022393	0.034753
	sd (MSE.PEp)	9.85E-07	1.17E-05	3.72E-05	0.000234	0.001305	0.003767	0.008976	0.011168	0.004036

Table A2. 2. Summary of MSEs for each bin according to different sample sizes (25, 50, 100 and 200) for equal haplotype frequency setting (0.25, 0.25, 0.25 and 0.25)

		MSE.PEs		MSE.PS		MSE.SEs	
Bin	Sample size	mean	std dev	mean	std dev	mean	std dev
1st	25	0.003645	0.000638	0.003645	0.000636	4.36E-06	1.92E-06
	50	0.001915	0.000364	0.001915	0.000363	2.19E-06	4.28E-07
	100	0.001005	0.000199	0.001005	0.000199	1.39E-06	2.32E-07
	200	0.000191	3.62E-05	0.000191	3.62E-05	4.24E-07	1.30E-07
5th	25	0.003563	0.000572	0.003562	0.000555	3.52E-05	5.82E-06
	50	0.001874	0.000326	0.001874	0.000318	1.87E-05	4.21E-06
	100	0.000973	0.000169	0.000971	0.000166	9.82E-06	3.06E-06
	200	0.000193	3.18E-05	0.00019	3.13E-05	4.17E-06	2.75E-06
10th	25	0.003685	0.00058	0.003657	0.000549	8.92E-05	2.16E-05
	50	0.001954	0.000343	0.001933	0.000325	5.46E-05	1.74E-05
	100	0.001017	0.000184	0.000997	0.000175	3.52E-05	1.38E-05
	200	0.000214	3.51E-05	0.000197	3.29E-05	1.97E-05	1.19E-05
25th	25	0.003882	0.000735	0.003618	0.00063	0.000401	0.000153
	50	0.00213	0.000375	0.001916	0.00031	0.000274	0.000151
	100	0.001144	0.00023	0.000965	0.000171	0.000227	0.00015
	200	0.000365	0.00014	0.000197	3.19E-05	0.000184	0.000146
50th	25	0.005442	0.001341	0.003641	0.000432	0.002222	0.00147
	50	0.003472	0.001278	0.001934	0.000228	0.001731	0.001328
	100	0.002371	0.001268	0.000955	0.000117	0.001516	0.001317
	200	0.001537	0.001319	0.000191	2.25E-05	0.001366	0.001318
75th	25	0.017761	0.00997	0.003689	0.00039	0.015144	0.010349
	50	0.015177	0.010261	0.001937	0.000236	0.013603	0.010566
	100	0.013789	0.010445	0.000977	0.000107	0.013057	0.010623
	200	0.012684	0.010825	0.000199	2.80E-05	0.012517	0.010903
90th	25	0.038661	0.01775	0.003562	0.000554	0.036935	0.020078
	50	0.037017	0.018479	0.001889	0.000331	0.035855	0.019788
	100	0.035889	0.019067	0.000982	0.000197	0.034988	0.019544
	200	0.035029	0.019472	0.000192	3.15E-05	0.034668	0.019345
95th	25	0.040782	0.007776	0.003061	0.000182	0.038705	0.009802
	50	0.03965	0.0081	0.001633	0.000132	0.038151	0.009428
	100	0.038662	0.008503	0.000812	7.23E-05	0.037828	0.009019
	200	0.037956	0.009043	0.000173	1.46E-05	0.037404	0.009178
99th	25	0.048261	0.007092	0.002978	0.00024	0.046918	0.008207
	50	0.053658	0.005134	0.001537	8.94E-05	0.053202	0.006229
	100	0.054178	0.004637	0.000789	7.23E-05	0.053724	0.00501
	200	0.054071	0.004503	0.000159	8.54E-06	0.053483	0.00455

Table A2. 3. Summary of MSEs for each bin according to different sample sizes (25, 50, 100 and 200) for unequal haplotype frequency setting (0.1, 0.2, 0.3 and 0.4)

		MSE.PEs		MSE.PS		MSE.SEs	
Bin	Sample size	mean	std dev	mean	std dev	mean	std dev
1st	25	0.003457	0.000678	0.003454	0.000674	1.49E-05	5.07E-06
	50	0.00182	0.000353	0.001816	0.00035	7.95E-06	2.59E-06
	100	0.000935	0.000186	0.000933	0.000184	4.30E-06	1.44E-06
	200	0.000174	3.21E-05	0.000173	3.22E-05	1.82E-06	1.04E-06
5th	25	0.003453	0.000834	0.00342	0.000802	7.46E-05	2.83E-05
	50	0.001801	0.000442	0.001786	0.000431	4.29E-05	1.93E-05
	100	0.000927	0.000221	0.000911	0.000216	2.72E-05	1.45E-05
	200	0.00019	4.01E-05	0.000177	4.04E-05	1.54E-05	1.18E-05
10th	25	0.003606	0.000855	0.003525	0.000798	0.000146	6.42E-05
	50	0.00188	0.000442	0.001827	0.000417	8.89E-05	4.99E-05
	100	0.000982	0.000242	0.000936	0.000225	6.42E-05	4.32E-05
	200	0.000218	6.05E-05	0.000179	4.28E-05	4.11E-05	3.71E-05
25th	25	0.003807	0.000936	0.00345	0.000839	0.000535	0.000303
	50	0.002118	0.000518	0.001811	0.000434	0.000387	0.000268
	100	0.0012	0.000334	0.00092	0.000216	0.000331	0.000259
	200	0.000444	0.000245	0.000177	4.29E-05	0.000277	0.000243
50th	25	0.00567	0.001622	0.003523	0.000728	0.002539	0.001423
	50	0.00368	0.001381	0.001829	0.00036	0.002093	0.001391
	100	0.002645	0.001333	0.000929	0.000176	0.001845	0.001362
	200	0.001829	0.001335	0.00018	3.82E-05	0.001692	0.001345
75th	25	0.009724	0.003587	0.003761	0.000578	0.006356	0.003872
	50	0.007083	0.003482	0.001955	0.000296	0.00536	0.003743
	100	0.005681	0.003619	0.000982	0.000136	0.004768	0.003727
	200	0.004561	0.003675	0.000193	3.10E-05	0.004337	0.00368
90th	25	0.020858	0.008114	0.003448	0.000469	0.018694	0.008883
	50	0.018491	0.008411	0.001813	0.000288	0.017435	0.008777
	100	0.017292	0.00874	0.000919	0.000117	0.016832	0.00887
	200	0.016436	0.008878	0.000179	2.49E-05	0.016388	0.008945
95th	25	0.026819	0.00946	0.003587	0.00053	0.024384	0.011243
	50	0.024342	0.010314	0.001769	0.000216	0.023088	0.011082
	100	0.023422	0.010759	0.000941	0.000103	0.022531	0.011095
	200	0.022557	0.011114	0.000185	2.74E-05	0.022386	0.011115
99th	25	0.037177	0.004117	0.00326	0.000135	0.036835	0.004711
	50	0.03583	0.004143	0.001646	4.62E-05	0.035379	0.004163
	100	0.035161	0.004238	0.000869	3.76E-05	0.034743	0.004075
	200	0.034777	0.004049	0.000185	1.72E-05	0.034657	0.003962