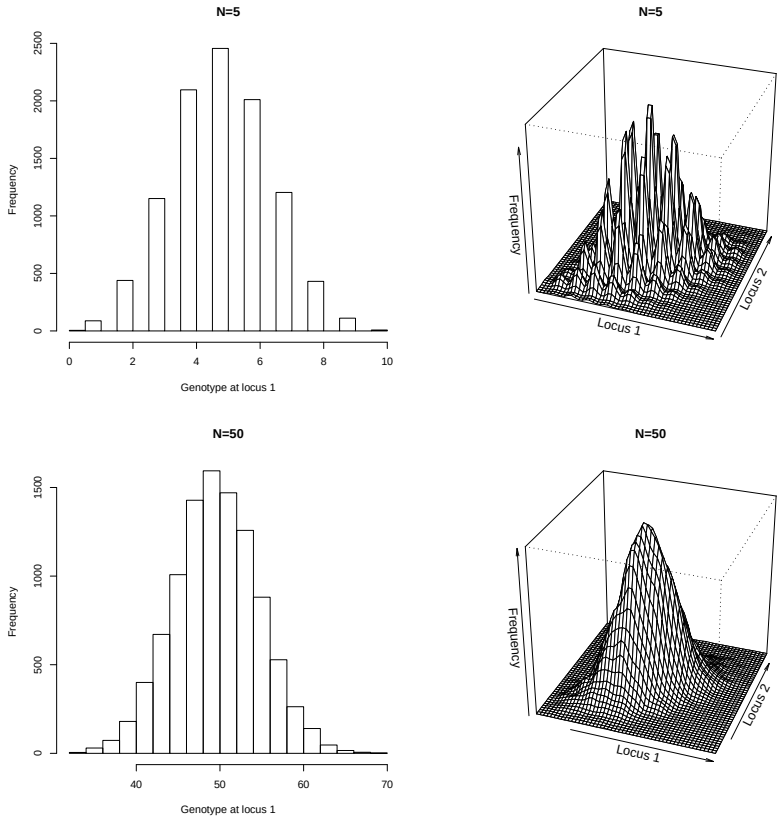
# Supplementary material to "PoooL: An efficient method for estimating haplotypes from large DNA pools"

## 1 About normality assumption for pooled genotypes

We approximate the distribution of pooled genotypes by normal distribution in our study. Since pool of $N$ individuals can be viewed as summation of $2N$ independent haplotypes, from central limit theorem, the asymptotic distribution of pooled genotypes are multivariate normal distribution. We illustrate this by Figure 1, from which we can see that the marginal distribution or the joint distribution of pooled genotypes at two loci can be well approximated by normal distribution, even when the pool size $N$ is not very large ($N = 5$).

Figure 1: Histograms of pooled genotypes of $N$ individuals (left column: genotypes at one locus; right panel: joint distribution at two loci)

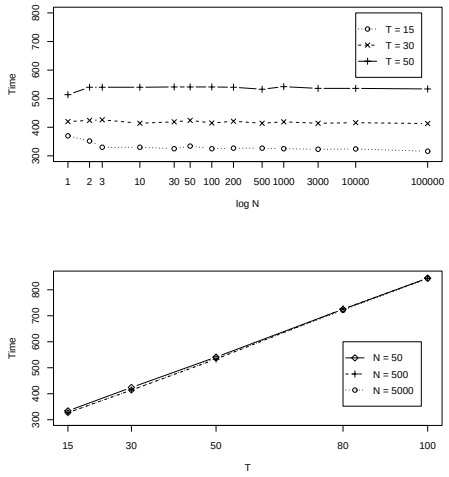



## 2 Computing efficiency of PoooL

The pool size is irrelevant in our algorithm and it is order $N!$ more efficient than the classical EM algorithm for large pools. Figure 2 illustrates this clearly. The computing expense is measured in seconds for estimating $q$-locus haplotype frequencies from data sets with $T$ pools each with size

$N$. In this figure, we use the Jain's data ($q = 3$) and perform 500 bootstrap resampling replicates (for estimating variances of the haplotype frequency estimates from PoooL) in 2000 simulation replicates. Clearly, from the figure the computing time is not affected by pool size $N$ and is linearly dependent on the number of pools.

Figure 2: Computing time (in seconds) of 2000 replicates of simulations (with 500 resamplings for each simulation) by using the PoooL program on an Intel(R) 2.2GHz notebook PC with 2 GB RAM.




## 3 More examples of population stratification

**Example 1.** In Table 1, population $A$ and $B$ have the same allele frequencies (0.4 and 0.3) on the two loci but they have different haplotype frequencies. Population $A$ is the major population, a small proportion, $\tau_B$, of individuals from population $B$ are mixed in forming DNA pools. It can be seen that the haplotype frequency estimates are biased due to the admixture of population $B$, but the biases are generally small when $\tau_B$ is small. $T = 30$, $N = 50$.

Table 1: Two-locus case: Haplotype frequency estimates when a proportion, $\tau_B$, of individuals from population $B$ is admixed in population $A$.

| SNPs | $\mathbf{p}$ | | $\hat{\mathbf{p}}$ | | | |
|---|---|---|---|---|---|---|
| 1 2 | $A$ | $B$ | $\tau_B = 0$ | $\tau_B = 0.1$ | $\tau_B = 0.2$ | $\tau_B = 0.3$ |
| 0 0 | 0.520 | 0.600 | 0.519 | 0.526 | 0.535 | 0.542 |
| 0 1 | 0.080 | 0.000 | 0.082 | 0.074 | 0.066 | 0.059 |
| 1 0 | 0.180 | 0.100 | 0.180 | 0.173 | 0.163 | 0.156 |
| 1 1 | 0.220 | 0.300 | 0.219 | 0.227 | 0.236 | 0.243 |
| $D'$ | 0.556 | 1.000 | 0.547 | 0.590 | 0.636 | 0.675 |

**Example 2.** Table 2 illustrates a 3-locus case of population admixture ($T = 30$, $N = 5$). The two populations share the same allele frequencies on the 3 loci (minor allele frequencies are 0.47, 0.11,

Table 2: Haplotype frequency estimates when a proportion, $\tau_B$, of individuals from population $B$ is admixed in population $A$.

| SNPs | $\mathbf{p}$ | | $\hat{\mathbf{p}}$ | | |
|---|---|---|---|---|---|
| 1 2 3 | $A$ | $B$ | $\tau_B = 0$ | $\tau_B = 0.1$ | $\tau_B = 0.2$ |
| 0 0 0 | 0.000 | 0.047 | 0.014 (0.017) | 0.017 (0.019) | 0.021 (0.020) |
| 1 0 0 | 0.082 | 0.010 | 0.072 (0.028) | 0.067 (0.034) | 0.058 (0.027) |
| 0 1 0 | 0.000 | 0.024 | 0.000 (0.001) | 0.001 (0.002) | 0.001 (0.002) |
| 1 1 0 | 0.000 | 0.001 | 0.011 (0.014) | 0.011 (0.014) | 0.011 (0.019) |
| 0 0 1 | 0.525 | 0.440 | 0.505 (0.043) | 0.497 (0.043) | 0.490 (0.042) |
| 1 0 1 | 0.283 | 0.392 | 0.285 (0.045) | 0.296 (0.045) | 0.311 (0.043) |
| 0 1 1 | 0.004 | 0.019 | 0.017 (0.020) | 0.019 (0.020) | 0.021 (0.022) |
| 1 1 1 | 0.106 | 0.067 | 0.097 (0.030) | 0.092 (0.030) | 0.088 (0.030) |

0.08) but have different haplotype frequencies (columns 3-4). Population $A$ is generated from data from Table 3 in the manuscript (Jain's data) and is of major interest. Population $B$ is noise. The biases are small when proportion of population $B$ is small ($\tau_B = 0.1, 0.2$). When $\tau_B = 0.1$, the maximum bias is 0.028.

**Example 3.** Table 3 gives another 3-locus example of population admixture ($T = 30$, $N = 5$). The two populations share the same allele frequencies on the 3 loci (minor allele frequencies are 0.47, 0.11, 0.08) but have different haplotype frequencies (columns 3-4). Again the admixture causes biases in the estimates and the biases are small when proportion of population $B$ is small ($\tau_B = 0.1, 0.2$). When $\tau_B = 0.1$, the maximum bias is 0.030.

Table 3: Haplotype frequency estimates when a proportion, $\tau_B$, of individuals from population $B$ is admixed in population $A$.

| SNPs | $\mathbf{p}$ | | $\hat{\mathbf{p}}$ | | |
|---|---|---|---|---|---|
| 1 2 3 | $A$ | $B$ | $\tau_B = 0$ | $\tau_B = 0.1$ | $\tau_B = 0.2$ |
| 0 0 0 | 0.000 | 0.050 | 0.013 (0.018) | 0.015 (0.018) | 0.019 (0.019) |
| 1 0 0 | 0.082 | 0.022 | 0.074 (0.029) | 0.068 (0.027) | 0.064 (0.027) |
| 0 1 0 | 0.000 | 0.000 | 0.000 (0.001) | 0.000 (0.001) | 0.001 (0.002) |
| 1 1 0 | 0.000 | 0.009 | 0.011 (0.016) | 0.010 (0.013) | 0.009 (0.013) |
| 0 0 1 | 0.525 | 0.390 | 0.503 (0.043) | 0.495 (0.042) | 0.485 (0.045) |
| 1 0 1 | 0.283 | 0.427 | 0.286 (0.043) | 0.301 (0.042) | 0.311 (0.044) |
| 0 1 1 | 0.004 | 0.090 | 0.018 (0.021) | 0.024 (0.023) | 0.029 (0.025) |
| 1 1 1 | 0.106 | 0.012 | 0.095 (0.029) | 0.087 (0.029) | 0.082 (0.032) |