Combining markers into haplotypes can improve population structure inference

Lucie M. Gattepaille\textsuperscript{a} and Mattias Jakobsson\textsuperscript{a,b,*}

\textsuperscript{a}Department of Evolutionary Biology, EBC, Uppsala University, Uppsala, Sweden
\textsuperscript{b}Science for Life Laboratory, Uppsala University, Uppsala, Sweden

August 19, 2011

ABSTRACT

High through-put genotyping and sequencing technologies can generate dense sets of genetic markers for large numbers of individuals. For most species, these data will contain many markers in linkage disequilibrium (LD). To utilize such data for population structure inference, we investigate the use of haplotypes constructed by combining the alleles at Single Nucleotide Polymorphisms (SNPs). We introduce a statistic derived from information theory, the \textit{Gain of Informativeness for Assignment} (GIA), which quantifies the additional information for assigning individuals to populations using haplotype data compared to using individual loci separately. Using a two-loci-two-allele model, we demonstrate that combining markers in linkage equilibrium into haplotypes always lead to non-positive GIA suggesting that combining the two markers is not advantageous for ancestry inference. However, for loci in LD, GIA is often positive suggesting that assignment can be improved by combining markers into haplotypes. Using GIA as a criterion for combining markers into haplotypes, we demonstrate for simulated data, a significant improvement of assigning individuals to candidate populations. For the many cases that we investigate, incorrect assignment was reduced between 26\% and 97\% using haplotype data. For empirical data from French, German and Swiss individuals, we can assign every individual to the correct population using haplotype data, whereas assignment is difficult with SNPs. Our results can be useful for challenging population structure and assignment problems, in particular for studies where large-scale population-genomic data is available.

INTRODUCTION

Structure of populations and assigning individuals to populations have attracted considerable attention in population genetics, conservation biology and ecology (PRITCHARD et al. 2000; BEAUMONT 2004; MANEL et al. 2005; PLATT et al. 2010). Since the introduction of Wright’s $F_{ST}$ (WRIGHT 1921, 1943), numerous studies of population structure have been conducted for a multitude of species using a variety of genetic or phenotypic markers. The recent development of high-throughput genotyping and sequencing technologies have resulted in a substantial increase in studies of population structure that are based on a large number of markers (e.g. JAKOBSSON et al. 2008; VONHOLDT et al. 2010; PLATT et al. 2010). At the same time, powerful clustering methods have been developed to infer population structure based on multi-loci genetic data (e.g. PRITCHARD et al. 2000; DAWSON and BELKHIR 2001; CORANDER et al. 2003; FRANCOIS et al. 2006; HUELSENBECK and ANDOLFATTO 2007; ALEXANDER et al. 2009). For most species, individuals rarely reproduce at random and this can create genetically differentiated subgroups within a population or species. Geographic barriers such as mountains, rivers and oceans can furthermore hinder random mating, thereby causing populations to be structured (HALE et al. 2001; ROSENBERG et al. 2005). In humans, cultural differences, such as language or religious beliefs, may play an additional role in shaping structure among individuals (Cavlalli-Sforza and Feldman 2003; BRYC et al. 2010; BEHAR et al. 2010). Large efforts have been made to characterize population structure, both at the global level (e.g. ROSENBERG et al. 2002; JAKOBSSON et al. 2008; LI et al. 2008) and at smaller scales (e.g. ROSENBERG et al. 2006; WANG et al. 2007; SEGUREL et al. 2008; NOVEMBRE et al. 2008; FRIEDLAENDER et al. 2008; TISHKOFF et al. 2009; REICH et al. 2009). Though population structure can give important information on the demographic history of a species and may lead to better understanding of evolutionary processes, population structure may also complicate certain investigations. For example, cryptic
population structure can lead to false positives in association studies (Marchini et al. 2004). Another problem may arise in forensics; if a suspect originates from a population that is genetically differentiated from the reference population, the difference in allele frequencies may lead to incorrect conclusions about matching DNA evidence to a suspect (Aitken and Taroni 2004; Weir 1996; Balding and Nichols 1994).

Assignment methods, in contrast to clustering methods, use prior knowledge about candidate groups in addition to genetic data to assign individuals of unknown origin to groups (Paetkau et al. 1995; Manel et al. 2005). These methods have been extensively used for conservation management (see e.g. Wasser et al. 2004; Gaskin et al. 2009) and parentage analysis (see e.g. Nielsen et al. 2001). Methods that focus on finding potential hybrids of particular types (e.g. first generation offspring and back-crosses) have also been developed (Anderson and Thompson 2002) and used for identifying hybrids between closely related species (Adams et al. 2007).

High-throughput sequencing and genotyping methods have generated dense sets of Single Nucleotide Polymorphisms (SNPs) for large samples of individuals for several organisms. Linkage Disequilibrium (LD) is strong for many SNPs in these dense sets (for most species), and these SNPs are therefore not independent markers. To overcome the problem of LD, some studies prune the set of SNPs before inferring population structure (e.g. Novembre et al. 2008; Bryc et al. 2010) and some studies analyze subsets of markers and combine the results for different subsets (Jakobsson et al. 2008). These approaches of overcoming the problem caused by closely linked markers do not take full advantage of all the information provided by the large number of SNPs. Instead, it may be possible to combine SNPs into haplotypes, which may integrate extra information about ancestry, potentially from recombination events that should in principle harbor information about ancestry similar to mutation events. A previous study has utilized haplotypes for revealing population structure, which point at somewhat different inference of population structure for SNPs and haplotypes (Jakobsson et al. 2008). Using simulations, Morin et al. (2009) demonstrated greater power of population structure inference using haplotypes in many, but not all, cases. However, it is unclear if, and under which conditions, haplotypes can be more powerful than single SNPs for inferring population structure or assigning individuals to populations.

In this paper, we first investigate if haplotype data can increase the statistical power of assigning individuals to populations compared to SNP data. Second, using a newly developed statistic, the Gain of Informativeness for Assignment (GIA), we characterize under which circumstances it may be advantageous to use haplotypes compared to using SNPs for ancestry inference. Third, we demonstrate by simulations and by using empirical SNP data from Europeans that assignment of individuals significantly improves through combining SNPs into haplotypes guided by GIA.

**THEORY**

We define a “haplotype-locus” as the combination of more than one SNP locus. The SNP loci in a haplotype-locus are not required to be consecutive along the chromosome. We define a “haplotype-allele” as a particular combination of alleles at the SNP loci constituting the haplotype-locus. For instance, for a haplotype-locus formed by $x$ SNPs, $2^x$ distinct alleles can exist, but the number of observed haplotype-alleles is typically much smaller than $2^x$ if $x$ is reasonably large. In addition, the number of distinct haplotype-alleles is upwardly bounded by the sample size.

In order to develop a statistic that quantifies under which circumstances it is advantageous for ancestry inference to combine markers into haplotype-loci, we start by considering a model of two multi-allelic loci denoted locus A and locus B. The combination of the two loci into a haplotype-locus is denoted locus H, and the possible haplotype-alleles are the combinations of alleles from locus A and locus B (see figure 1 for notation). Note that this model can be generalized to handle any number of markers by recursively merging two loci into one multi-allelic haplotype-locus. Locus A and B may be in LD, which can, for example, be quantified with the $D$ statistic (Lewontin and Kojima 1960). We consider $K$ randomly mating populations and we assume that the allele frequencies at each locus in each population are known.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$a_1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$a_2$</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$b_1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_2$</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>$x_{11}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$x_{12}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$x_{21}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$x_{22}$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Notation for frequencies of the two alleles at locus A, the two alleles at locus B, and the four alleles at haplotype-locus H formed by combining the alleles at locus A and locus B.

Rosenberg et al. (2003) derived a criterion based on information theory to evaluate the efficiency of a marker for assigning individuals to one of $K$ populations. This criterion, the Informativeness for Assignment (IA) can be computed for bi- or multi-allelic loci, such as SNPs, microsatellites or haplotype-loci,

$$IA = \sum_{j=1}^{N} \left( -\overline{p}_j \log \overline{p}_j + \sum_{i=1}^{K} \frac{p_j^{(i)}}{K} \log p_j^{(i)} \right),$$

where $N$ is the number of alleles for the locus, $K$ is the number of populations, $p_j^{(i)}$ is the frequency of allele $j$ in population $i$. 


and \( p_j \) is the average frequency of allele \( j \) across populations,

\[
\bar{p}_j = \frac{1}{K} \sum_{i=1}^{K} p_{j}^{(i)}.
\]

Using the \( IA \) statistic, we define the Gain of Informativeness for Assignment (GIA) as

\[
GIA = IA(H) - [IA(A) + IA(B)],
\]

where \( IA(H) \) is the Informativeness for Assignment of the haplotype-locus and \( IA(A) \) and \( IA(B) \) are the informativeness of locus A and locus B respectively. Since \( IA \) is non-negative and bounded upward by \( \log K \), \( GIA \) is restricted to \([-2 \log K, \log K]\).

By comparing the information content about ancestry of the haplotype to the sum of the information content of each marker, \( GIA \) is specifically designed to answer the question of whether two markers can improve the power of assigning individuals to candidate populations by combining the markers into a haplotype-locus. As can be seen from equations 1 and 2, to compute \( GIA \), we need to know the allele frequencies of the two loci and the allele frequencies of the haplotype-locus. When addressing assignment problems, phased data from candidate populations can typically be used to estimate the SNP- and haplotype-allele frequencies, followed by the use of \( GIA \) to determine which loci to combine to haplotype-loci for optimal power. Guided by this information, individuals of unknown origin could then be assigned to candidate populations on the basis of haplotype data (see the results-section for explicit examples of this procedure).

\( GIA \) is not a simple function of the allele frequencies and the haplotype-allele frequencies. For example, the sign of \( GIA \) cannot be determined by a simple rule of thumb based on allele frequencies. However, for the special case of bi-allelic markers, we can show that when two loci are in linkage equilibrium, \( GIA \leq 0 \). In order to arrive at that result, we note that because the loci are bi-allelic, only the frequencies of one allele for each locus are needed to characterize \( GIA \). Recall also that \( D \) can be defined as the difference between the frequency of a haplotype-allele and the product of the frequencies of its constitutive alleles so that haplotype-allele frequencies in equation 2 can be replaced by \( D \) and allele frequencies (e.g. \( x_{11} = a_1 b_1 + D \)).

**Theorem.** Let A and B be two bi-allelic loci and \( H \) be their associated haplotype-locus. Consider \( K \) randomly mating populations. For population \( i \), let \( a_i^{(1)} \) and \( b_i^{(1)} \) be the allele frequencies at locus A and locus B respectively. Then, for all the frequency distributions of the alleles,

\[
\forall i \in 1 \ldots K, \; D_i = 0 \Rightarrow GIA = IA(H) - [IA(A) + IA(B)] \leq 0
\]

with equality if and only if \( \sum_{i=1}^{K} \sum_{k=1}^{i} (a_i^{(1)} - a_i^{(k)}) (b_i^{(1)} - b_i^{(k)}) = 0 \).

A proof of the theorem is given in the appendix. This theorem demonstrates that when locus A and locus B are in linkage equilibrium within all populations, the haplotype-locus \( H \) provides less information (or the same amount) for assigning individuals to populations than locus A and locus B provides when used separately. Intuitively, since there is no correlation between the allele frequencies at locus A and the allele frequencies at locus B, we expect the combination of alleles into haplotype-alleles to arise randomly within each population.

**GIA for two populations:** We study equation 2 for the two population case \( (K = 2) \) and for two bi-allelic markers. To reduce the complexity of the problem, we assume that the level of LD is dominated by linkage of the two markers, and that the two populations have similar demographic histories, so that \( D_1 = D_2 = D \). Five parameters characterize our problem: \( D, \; a_1^{(1)}, \; a_1^{(2)}, \; a_2^{(1)} \) and \( a_2^{(2)} \). The haplotype-allele frequencies must be greater than or equal to zero in both populations, which limits the range of \( D \) and the range of the allele frequencies at locus A and locus B, constraints are summarized in Table 1. As an example, we study the behavior of \( GIA \) as a function of \( a_1^{(1)} \) and \( a_1^{(2)} \) for \( D = 0.1 \) and different fixed values of \( b_1^{(1)} \) and \( b_1^{(2)} \). Figure 2 shows that \( GIA \) is positive for some parts of the parameter space, but it can also be negative, depending on the values \( a_1^{(1)}, \; a_1^{(2)}, \; b_1^{(1)} \) and \( b_1^{(2)} \). Figure 2A shows the values of \( GIA \) when \( b_1^{(1)} = b_1^{(2)} = 0.2 \) and \( D = 0.1 \) for the entire range of possible values of \( a_1^{(1)} \) and \( a_1^{(2)} \), a case in which locus B is uninformative on its own \( (IA(B) = 0) \) since it has identical allele frequencies in both populations. \( GIA \) is non-negative for all possible values of \( a_1^{(1)} \) and \( a_1^{(2)} \), which means that the haplotype-locus contains more information for assigning individuals to populations than the two loci used separately. The intuition behind this result is that locus A has only two alleles, whereas the haplotype-locus can have up to four different alleles, increasing the possibility for the haplotype-alleles to uniquely characterize populations, which makes the assignment of individuals easier.

Figures 2B and 2C show that the sign and magnitude of \( GIA \) varies depending on the values of the allele frequencies at locus A. The borders of the surfaces are defined by the constraints on \( a_1^{(1)} \) and \( a_1^{(2)} \) given in table 1 and at each border of the surfaces, at least one haplotype-allele frequency equals zero in one of the two populations, i.e., private for one population. There are two interesting points on the surfaces, the leftmost tip and the rightmost tip. Although they share the same property of being the only cases where two haplotype-alleles are private, the rightmost tip yields the maximum \( GIA \) whereas the leftmost tip yields a negative \( GIA \). The absolute difference \( |a_1^{(1)} - a_1^{(2)}| \) distinguishes the two points, which is greater for the leftmost tip, resulting in a greater \( IA(A) \) and therefore a smaller \( GIA \) than for the rightmost tip. Nevertheless, they are both local maxima, which is caused by the often substantial informativeness of private alleles.
The exponent $D$ entire range (figure 3D), or change sign depending on
Figure 2:
Table 1: Constraints on $D$ and allele frequencies for locus A and locus B to ensure admissibility of haplotype-allele frequencies in population $i$. The exponent $(i)$ is omitted for convenience.

![Figure 2](image1.png)

Figure 2: $GIA$ as a function of $a_1^{(1)}$, $a_1^{(2)}$, when $D = 0.1$, for different fixed values of $b_1^{(1)}$ and $b_1^{(2)}$. (A) $b_1^{(1)} = 0.2$ and $b_1^{(2)} = 0.2$; (B) $b_1^{(1)} = 0.3$ and $b_1^{(2)} = 0.6$; and (C) $b_1^{(1)} = 0.15$ and $b_1^{(2)} = 0.6$.

We also investigate the behavior of $GIA$ as a function of $D$ when all the allele frequencies are fixed and $GIA$ is therefore completely determined by IA(H). Figure 3 shows four examples of $GIA$ as functions of $D$, across the range of possible values of $D$, for different values of $a_1^{(1)}$, $a_1^{(2)}$, $b_1^{(1)}$ and $b_1^{(2)}$. We first observe that if $D = 0$, $GIA \leq 0$ (consistent with the theorem). For $a_1^{(1)} = 0.4$, $a_1^{(1)} = 0.3$ and $b_1^{(1)} = b_1^{(2)} = 0.2$ (figure 3A), $GIA$ is non-negative for the whole range of $D$. This example is similar to the example in figure 2A, for which locus B was also uninformative.

The sign and the magnitude of $GIA$ varies as a function of $D$ for fixed allele frequencies of locus A and locus B. $GIA$ can be positive for the entire range of $D$ (figure 3A), negative for the entire range (figure 3D), or change sign depending on $D$ (figures 3B and 3C). The range of $D$ is defined by the constraints that all haplotype-allele frequencies have to be non-negative. The two extreme values for each case in figure 3 correspond to one of the eight haplotype-allele frequencies (four haplotype-allele frequencies in each population) being equal to zero in one population, which means being a private allele for the other population.

In summary, although there are a number of predictable behaviors of $GIA$ – such as that $GIA \leq 0$ when markers are in linkage equilibrium and that $GIA$ is often large for cases where private alleles exist – $GIA$ is not a trivial function of LD or allele frequencies.

![Figure 3](image2.png)

Figure 3: $GIA$ as a function of $D$ for fixed values of the allele frequencies in both populations. (A) $a_1^{(1)} = 0.4$, $a_1^{(2)} = 0.3$, and $b_1^{(1)} = b_1^{(2)} = 0.2$; (B) $a_1^{(1)} = 0.2$, $a_1^{(2)} = 0.3$, $b_1^{(1)} = 0.3$, and $b_1^{(2)} = 0.6$; (C) $a_1^{(1)} = 0.4$, $a_1^{(2)} = 0.3$, $b_1^{(1)} = 0.2$, and $b_1^{(2)} = 0.5$; and (D) $a_1^{(1)} = 0.15$, $a_1^{(2)} = 0.8$, $b_1^{(1)} = 0.2$, and $b_1^{(2)} = 0.8$. 

| Sign of $D$ | on $D$ | on $b_1|D$ | on $a_1|D, b_1$ |
|------------|--------|-----------|----------------|
| positive   | $D \leq \frac{1}{4}$ | $|b_1 - \frac{1}{2}| \leq \sqrt{1 - 4D}$ | $\frac{D}{1 - b_1} \leq a_1 \leq 1 - \frac{D}{b_1}$ |
| negative   | $D \geq -\frac{1}{4}$ | $|b_1 - \frac{1}{2}| \leq \sqrt{1 + 4D}$ | $-\frac{D}{b_1} \leq a_1 \leq 1 + \frac{D}{b_1}$ |
RESULTS

Comparing GIA and performance of assignment: To assess how haplotype-loci that are constructed based on GIA perform for assigning individuals to populations, we evaluate assignment in a two-population case for a wide range of allele frequencies and levels of linkage disequilibrium. We investigate a case of 200 haploid individuals, 100 individuals from each population, where each individual is assumed to be typed for 40 pairs of SNPs. We generate a discrete set of haploid gene-copies (for a pair of SNPs) for each population that satisfy a particular choice of allele frequencies and levels of LD (see table 2). This set of gene-copies are randomly permuted to generate a set of 40 pairs of SNPs, which ensures that the pairs of SNPs are independent of each other (conditional on the allele frequencies). This procedure guarantees that all the SNP pairs have the same allele frequencies for SNP A, SNP B, the A-B haplotype-locus, and consequently the same level of LD between the two SNPs. Note that within a population, most of the LD in the sample is a result of the linkage between the two SNPs in each pair.

For these population-genetic data, we use the software STRUCTURE (PRITCHARD et al. 2000; FALUSH et al. 2003), to assign the 200 haploid individuals to two clusters (no-admixture model, burn-in period of 20,000 iterations followed by 5,000 iterations from which estimates were obtained), using either the 80 SNPs or the 40 haplotype-loci obtained by combining each pair of SNPs into one haplotype-locus. From the STRUCTURE result, the Mean Incorrect Assignment Proportion (MIAP) is computed, which is the average proportion of individuals that are assigned to the incorrect population. For a given set of allele frequencies, we generate 100 different replicate samples using the data-randomization procedure described above, assign individuals to populations, and compute the average (across replicates) of MIAP. For comparison, $F_{ST}$ values for the SNP pairs, as well as $F_{ST}$ values for the haplotype-loci are computed. Similarly to IA, $F_{ST}$ also relies on information about allele frequencies.

| $a_1^{(1)}$ | $a_1^{(2)}$ | $b_1^{(1)}$ | $b_1^{(2)}$ | $|D|$ | $\tau^2$ | GIA | MIAP SNPs | MIAP Hapl. | $F_{ST}(\text{SNPs})$ | $F_{ST}(\text{Hapl.})$ |
|-----------|-----------|-----------|-----------|------|--------|-----|--------|-----------|----------------|----------------|
| 0.41 0.60 | 0.17 0.05 | 1.5×10^{-4} | 1.3×10^{-6} | $-6.29×10^{-4}$ | 0.0976 | 0.1209 | 0.0606 | 0.0498 |
| 0.62 0.81 | 0.38 0.25 | 0.0384 | 0.0500 | 1.59×10^{-2} | 0.1212 | 0.0719 | 0.0517 | 0.0844 |
| 0.38 0.17 | 0.11 0.15 | 0.0413 | 0.0998 | 1.03×10^{-2} | 0.1642 | 0.1140 | 0.0612 | 0.0689 |
| 0.47 0.32 | 0.21 0.12 | 0.0685 | 0.1500 | 1.07×10^{-2} | 0.4286 | 0.1519 | 0.0301 | 0.0589 |
| 0.11 0.23 | 0.26 0.18 | 0.0514 | 0.1846 | 7.55×10^{-2} | 0.4186 | 0.0015 | 0.0229 | 0.0568 |
| 0.05 0.01 | 0.13 0.05 | 0.0215 | 0.2004 | $-3.23×10^{-3}$ | 0.2422 | 0.2625 | 0.0257 | 0.0226 |
| 0.61 0.88 | 0.15 0.03 | 0.0589 | 0.2514 | $-2.19×10^{-2}$ | 0.0372 | 0.0460 | 0.1400 | 0.1168 |
| 0.31 0.35 | 0.23 0.30 | 0.1018 | 0.2981 | 3.01×10^{-2} | 0.4736 | 0.1376 | -0.0022 | 0.0260 |
| 0.08 0.21 | 0.03 0.11 | 0.0522 | 0.3599 | $-8.53×10^{-3}$ | 0.1740 | 0.1988 | 0.0503 | 0.0419 |
| 0.38 0.17 | 0.11 0.23 | 0.0895 | 0.3659 | 5.78×10^{-2} | 0.0444 | 0.0047 | 0.0731 | 0.0901 |
| 0.04 0.08 | 0.05 0.08 | 0.0222 | 0.3996 | 5.33×10^{-2} | 0.3738 | 0.0353 | 6.17×10^{-4} | 0.0527 |
| 0.71 0.61 | 0.60 0.49 | 0.1575 | 0.4560 | $-3.96×10^{-3}$ | 0.4205 | 0.4962 | 0.0133 | 0.0081 |
| 0.28 0.14 | 0.25 0.22 | 0.1196 | 0.4974 | 8.11×10^{-3} | 0.3747 | 0.2859 | 0.0194 | 0.0133 |
| 0.05 0.01 | 0.11 0.01 | 0.0172 | 0.5645 | 6.02×10^{-3} | 0.1701 | 0.0768 | 0.0562 | 0.0687 |
| 0.18 0.36 | 0.28 0.37 | 0.1582 | 0.6043 | 2.12×10^{-2} | 0.2387 | 0.0851 | 0.0378 | 0.0460 |
| 0.17 0.37 | 0.13 0.25 | 0.1327 | 0.6486 | $-1.07×10^{-2}$ | 0.1516 | 0.1618 | 0.0653 | 0.0597 |
| 0.14 0.34 | 0.19 0.38 | 0.1521 | 0.6913 | $-1.75×10^{-2}$ | 0.1042 | 0.1251 | 0.0848 | 0.0733 |
| 0.01 0.06 | 0.99 0.89 | 0.0316 | 0.7582 | $-7.37×10^{-3}$ | 0.1759 | 0.1227 | 0.0576 | 0.0539 |
| 0.33 0.26 | 0.33 0.24 | 0.1843 | 0.8138 | $-2.45×10^{-3}$ | 0.4545 | 0.4920 | 0.0057 | 0.0044 |
| 0.08 0.10 | 0.08 0.14 | 0.0798 | 0.8413 | 9.66×10^{-3} | 0.4378 | 0.2871 | 0.0011 | 0.0040 |
| 0.06 0.08 | 0.06 0.10 | 0.0642 | 0.8913 | 4.35×10^{-3} | 0.4408 | 0.3645 | -0.0028 | -0.0020 |
| 0.28 0.07 | 0.28 0.07 | 0.1283 | 0.9516 | $-3.51×10^{-2}$ | 0.0483 | 0.0474 | 0.1332 | 0.1294 |
| 0.81 0.69 | 0.81 0.69 | 0.1839 | 1 | $-9.67×10^{-3}$ | 0.3272 | 0.4344 | 0.0280 | 0.0280 |

Table 2: The Mean Incorrect Assignment Proportion (MIAP) obtained by assigning 200 haploid individuals to either of two populations using STRUCTURE based on 80 SNPs, or based on 40 haplotype-loci, and for various allele frequencies and levels of LD. Values of $|D|$ and $\tau^2$ are means across populations. The values presented are averages across 100 replicate cases. GIA, $F_{ST}$ based on the 80 SNPs and $F_{ST}$ based on the 40 haplotype-loci are given for comparison. $F_{ST}$ values are computed using eqn. 5.3 in WEIR (1996). For MIAP, the smallest values between SNPs and haplotypes of incorrect assignments are highlighted in boldface.

Table 2 shows the performance of the assignment based on the 80 SNPs and based on the 40 haplotype-loci for various choices of allele frequencies and levels of LD. In most cases when GIA is positive, the MIAP values are lower for the haplotype-loci than for the SNPs. Similarly, when GIA is negative, the MIAP values are in most cases lower for the SNPs than for the haplotype-loci. For the choices of allele frequencies and levels of LD in table 2, figure 4 shows the difference between the MIAP based on SNPs and the MIAP based on haplotype-loci (i.e. improved assignment due to haplotype-
loci) as a function of GIA (figure 4A), the mean (across populations) of |D| (|D|, figure 4B), the mean (across populations) of $r^2$ ($r^2$, figure 4C), and the difference in $F_{ST}$ between the 40 haplotype-loci and the 80 SNPs (figure 4D). The improved assignment due to using haplotype-loci is positively correlated with GIA (Pearson: $\rho = 0.748$, $p = 4 \times 10^{-5}$), the correlation is non-significant with $|D|$ and $r^2$ ($\rho = -0.289$, $p = 0.16$ and $\rho = -0.302, p = 0.18$ respectively). The improved assignment is neither correlated with $F_{ST}$ for haplotype-loci nor $F_{ST}$ for SNPs ($\rho = -0.037, p = 0.87$ and $\rho = 0.401, p = 0.06$, respectively), but it is positively correlated with the difference between $F_{ST}$ for haplotype-loci and $F_{ST}$ for SNPs ($\rho = 0.790, p = 7 \times 10^{-6}$). GIA and the difference in $F_{ST}$ values appear to be good indicators of how assignment can be improved by combining SNPs into a haplotype-loci. The outlier observed far from the regression line in figure 4A corresponds to the 10th entry in table 2. For this set of allele frequencies, 40 pairs of SNPs is enough to obtain a very accurate assignment (MIAP close to 0) and there is not much room for improvement when combining the SNPs into haplotype-loci. GIA and the difference in $F_{ST}$ values are correlated ($\rho = 0.792$), suggesting that the two statistics contains similar information despite the fact that GIA is based on a measure of information whereas $F_{ST}$ measures differentiation, but there are similarities of the two statistics as well. Indeed, if the differentiation between the two populations is easier to capture when considering haplotype-loci compared to considering SNPs separately, we would expect that assignment also improves for haplotype data compared to SNP data.

Improving assignment using GIA – a simulation study:
For empirical population genetic data, allele frequencies and levels of LD vary extensively among loci. GIA is defined for multi-allelic markers and can be used for assessing the usefulness of combining not only pairs of SNPs, but also for combining haplotype-loci themselves. Thus, GIA can be used for large numbers of SNPs. In order to demonstrate the utility of GIA, we compare the results of the assignment of 200 haploid individuals originating from two populations and based on 1,000 SNPs using different strategies of dealing with the SNPs, e.g., by pruning the SNPs or combining them into haplotype-loci. We simulate the 200 haploid individuals with the software ms (Hudson 2002) from a two-island-model with migration rate $m$ (migrants/generation) and an effective population size of 1000. Each haploid individual represents a DNA fragment of 4.2 Mb with a total scaled recombination rate of $\rho = 4N\tau = 150$ or $\rho = 4N\tau = 1,500$ (where $N$ is the population size and $\tau$ is the recombination rate per generation for the entire fragment). We repeat the simulation 100 times for a given migration rate and a given recombination rate. For each sample, we assign the 200 individuals using STRUCTURE based on 7 different treatments of the SNPs:

a) Using all 1,000 SNPs.

b) Using a subset of the SNPs obtained by pruning. We prune the set of SNPs with the program PLINK (Purcell et al. 2007), to remove SNPs that are in high LD (rejection threshold of $r^2 = 0.1$, windows of 20 SNPs and shifts of 5 SNPs).

c) Combining the SNPs into haplotype-loci with a greedy algorithm that recursively combines the pair of loci that has the greatest GIA among all the pairwise comparisons of loci until no remaining pair of loci has a positive GIA. We refer to this strategy as MaxGIA.

d) Using a set of randomly formed haplotype-loci with a haplotype length distribution matching the haplotype length distribution of the set in c. We call this strategy RandomHaplotypes.

e) Using the set of SNPs and haplotype-loci obtained with the following algorithm: starting at the first SNP, if $GIA$
is positive between SNP 1 and SNP 2, combine them into a haplotype. Compute \( \text{GIA} \) for the SNP 1-SNP 2 haplotype and SNP 3, combine them into a haplotype if \( \text{GIA} \) is positive. Repeat this process until a SNP \( s \) is found for which the haplotype-locus and SNP \( s \) have a non-positive \( \text{GIA} \). Repeat the process starting from SNP \( s \). We refer to this strategy as \( \text{NeighborGIA} \).

f) Using a set of haplotype-loci formed by neighboring SNPs obtained by randomly permuting the break-points of the haplotype-loci set in \( e \), so that the haplotype length distribution is the same as in \( e \). We call this strategy \( \text{RandomNeighbor} \).

g) Combining the SNPs into haplotype-loci with a greedy algorithm that recursively combines the pair of loci that has the greatest \( \delta = F_{ST}(H) - F_{ST}(M1,M2) \) among all the pairwise comparisons of loci until no remaining pair of loci has a positive \( \delta \). \( F_{ST}(H) \) denotes \( F_{ST} \) for a haplotype-loci, and \( F_{ST}(M1,M2) \) denotes \( F_{ST} \) computed for the two markers constituting the haplotype-loci. We refer to this strategy as \( \text{MaxF}_{ST} \).

\[
\begin{align*}
\text{MIAP} &= \frac{1}{n} \sum_{i=1}^{n} \text{MIAP}_i \\
\text{F}_{ST} &= \frac{1}{n} \sum_{i=1}^{n} \text{F}_{ST,i}
\end{align*}
\]

For each sample, migration rate and strategy, we record the performance of assigning individuals to populations that is obtained from \( \text{STRUCTURE} \) (with the same settings as above). Figure 5 shows MIAP for the different strategies (no combination, pruning, \( \text{MaxGIA} \), \( \text{RandomHaplotypes} \), \( \text{NeighborGIA} \), \( \text{RandomNeighbor} \), and \( \text{MaxF}_{ST} \)) for a range of migration rates \( \rho \) and scaled recombination rates of \( \rho = 150 \) and \( \rho = 1,500 \). The \( \text{GIA} \) and the \( \text{F}_{ST} \) based strategies require some knowledge about allele frequencies for the considered markers, including the haplotype-loci formed in the iterative processes. In the context of an assignment problem, this information can be obtained from phased data for candidate populations. In this simulation study, we estimate the allele frequencies directly from the sample and use our knowledge of the individuals’ true ancestry. Thus, improvement based on the \( \text{GIA} \) or the \( \text{F}_{ST} \) strategies is to some degree magnified by the fact that we are using information about the individuals’ true ancestry to compute the allele frequencies. However, the \( \text{NeighborGIA} \) strategy uses the same information as the \( \text{MaxGIA} \) and the \( \text{MaxF}_{ST} \) strategies, and the improvement obtained for the \( \text{MaxGIA} \) and the \( \text{MaxF}_{ST} \) strategies cannot be explained solely by using information about the individuals’ ancestry.

For both recombination rates, the \( \text{MaxGIA} \) and the \( \text{MaxF}_{ST} \) strategies for combining SNPs show the least incorrect assignments, but recombination rate has a strong impact on the accuracy of the assignment. For the high-recombination case (\( \rho = 1,500 \)), the markers are less correlated and the set of markers carry more information about ancestry than the markers in the low-recombination case. Furthermore, as expected, when the migration rate increases (and \( \text{F}_{ST} \) decreases), MIAP also increases for all seven strategies. However, for the high-recombination case and a migration rate of 0.01, the \( \text{MaxGIA} \) and the \( \text{MaxF}_{ST} \) strategies can uncover the structure with (on average) less than 2% incorrect assignment compared to 37% using the full set or the pruned set of SNPs (figure 5B). Combining neighboring SNPs that have positive \( \text{GIA} \) also improves the assignment, but to a lesser extent than the \( \text{MaxGIA} \) strategy. For both choices of recombination rates, the strategies that combine SNPs into haplotypes in a random manner (\( \text{RandomHaplotypes} \) and \( \text{RandomNeighbor} \)) result in poor assignment. Thus, the improved assignment for \( \text{MaxGIA} \), and to some degree \( \text{NeighborGIA} \), compared to the pruning or no combination strategies are likely to be the result of using \( \text{GIA} \) as a criterion for combining SNPs into haplotypes and not just a result of randomly combining SNPs into haplotypes. However, for \( \rho = 1,500 \), the strategy \( \text{RandomNeighbor} \), which consists in randomly combining neighboring SNPs, increases the accuracy of the assignment compared to the pruning or no combination strategies. Finally,
In the case of 1% migrants per generation ($m = 0.01$, the greatest migration rate that we investigate), the distribution of MIAP for the 100 replicates varies depending on the strategy for treating the SNP data. Six distributions of MIAP (based on different treatments of the SNPs) for the low-recombination case ($\rho = 150$) are shown in figure 6 and the corresponding distributions of MIAP for the high-recombination case ($\rho = 1,500$) are shown in figure 7. For $\rho = 150$, the distribution of MIAP based on the MaxGIA strategy is spread over a range of values compared to the results of the other strategies, which are skewed towards 0.5, the expected value of MIAP for random assignment of individuals to populations (but note that this expected value may be slightly smaller for finite population sizes and unlabeled populations). So, as also shown by the mean MIAP in figure 5A, MaxGIA is the most accurate strategy, but there are also cases of poor assignment using this strategy. If we increase the recombination rate, all six distributions of MIAP move away from 0.5, except for RandomHaplotypes. The distributions of MIAP for RandomNeighbor, pruning or no combination strategies are similar and have large variances. The distribution of MIAP for the MaxGIA strategy is skewed towards 0, demonstrating superior assignment accuracy compared to the other strategies.

Figure 6: Histograms of the Mean Incorrect Assignment Probabilities (MIAP) for 100 replicates of simulated data from a two-island-model with migration rate $m$ equal to 0.01 and a scaled recombination rate of $\rho = 150$. The simulated SNP data is combined according to six different strategies, no combination, pruned set, MaxGIA, RandomHaplotypes, NeighborGIA and RandomNeighbor, and MIAP is computed for each strategy based on assignment of individuals using STRUCTURE.

In order to get an idea of how many SNPs make up the haplotype-loci that are constructed using the MaxGIA strategy, we compute the distribution of the number of SNPs in haplotype-loci for four different migration rates and for two different recombination rates (figure 8). All the length distributions show a clear mode, and the value of the mode appears to increase with increasing migration rate. This observation suggests that when it becomes more difficult to assign individuals to populations because of higher migration rate, longer haplotype-loci may increase the accuracy of the assignment. For the low-recombination case ($\rho = 150$), there is also a second mode at one single SNP (for all but the lowest migration rate), showing that many SNPs are not combined with other SNPs for these cases. In general however, the recombination rate appears to have little impact on the length distribution of the majority of haplotype-loci.

Improving assignment using GIA – POPRES data: To investigate if haplotype-loci can improve ancestry inference for empirical population genetic data, we use SNP-chip data from the POPRES panel that contain some 1385 individuals from Europe, (NELSON et al. 2008), which have been genotyped for some 500,000 SNPs. We phased all individuals using fastPHASE (SHEET and STEPHENS 2006), version 1.4 (‘haplotype clusters’ set to 20 and 20 runs of the EM algorithm), which generated “best guess” estimates of the phase of each of the two haploid copies for each individual.

We first conduct a cross validation study for the 89 French and the 71 German individuals in the POPRES collection (NELSON et al. 2008) and focus on the phased data of 5,637 SNPs on chromosome 22. To construct a training set, 45 French in-
individuals and 36 German individuals were randomly sampled, and the remaining 44 French and 35 German individuals make up the validation set. Using the MaxGLA strategy, we build a set of haplotype-loci using estimated allele frequencies from the training set of individuals. This set contains 460 haplotype-loci and the configuration of SNPs is known so that we can combine the SNPs in the validation set to make up the same haplotype-loci. We perform the assignment of the individuals in the validation set using STRUCTURE and using Principal Component Analysis (PCA), for either the entire set of SNPs or the set of haplotype-loci. For STRUCTURE, we compute MIAP (averaged across 10 replicate runs) for assigning the French and German individuals to candidate populations for both the training and the validation set.

Figure 8: Distribution of the length in number of SNPs of the haplotype-loci constructed with the MaxGLA strategy, computed for 100 replicate simulations and for four different migration rates. Results for two different recombination rates are presented; (A) a high-recombination case ($\rho = 150$) and (B) a low-recombination case ($\rho = 1,500$).

For assigning individuals in the training set using haplotype-loci, the assignment is highly accurate (no miss-classified individuals) in contrast to using SNPs (39% miss-classified individuals). However, assigning individuals in the validation set performs poorly for both haplotype-loci (44% miss-classified individuals) and SNPs (37% miss-classified individuals). For this particular case, it appears as assigning individuals to populations using STRUCTURE only works well based on haplotype data for the training set. However, PCA based on the haplotype data clusters the individuals in both the training set and the validation set (figure 9C-D), and individuals can be assigned to populations with high accuracy (in contrast to using SNPs, figure 9A-B). To perform the PCA, the haplotype data is transformed to a matrix of haplotype-alleles vs. chromosomes (two copies from each individual) where entries in the matrix denote presence (1), or absence (0), of a haplotype-allele in a particular chromosome. For both the training set and the validation set, the first component of such PCA based on haplotypes reveals a clear clustering of the chromosomes, according to French or German origin. The higher-order principal components all single out one chromosome from the other chromosomes (see e.g. PC2 in figure 9C-D).

Figure 9: Principal Component Analysis (PCA) for the individuals in the training set (A and C), for the individuals in the validation set (B and D), based on the 5,637 SNPs (A and B), and based on the 460 haplotype-loci constructed from the 5,637 SNPs (C and D). Each plot shows the two first PCs, and each individual is represented by 2 points in each plot, one point for each chromosome. The chromosomes from French individuals are represented by the red squares and the chromosomes from German individuals are represented by the blue triangles.

For investigating a more challenging and realistic application, we assign 209 individuals from Switzerland (84 Swiss-German and 125 Swiss-French) using a training set of 89 French and 71 German individuals from the POPRES data. We use the same procedure and the same 5,637 SNPs as for the cross-validation study above, and the haplotype-loci (in total 358) are
constructed using the MaxGIA strategy based on all the French and German individuals. For the haplotype-data, in contrast to the SNP data, the PCA clusters the Swiss individuals into two groups corresponding to Swiss-German and Swiss-French (figure 10B). The clustering based on haplotype data is also much more distinct compared to using all SNPs from all chromosomes; compare for example figure 10B with figure 1b in Novembre et al. (2008), which resembles our figure 10A. It might be tempting to conclude from this observation that in Switzerland, language is a barrier to geneflow – which may be the case –, but this conclusion does not follow from being able to assign individuals to two language-groups based on French and German candidate populations. For example, using the set of haplotype-loci constructed from the training set of French and German individuals, Irish and Portuguese individuals from POPRES can easily be distinguished based on the first PC from a PCA (data not shown). To conclude, by finding informative combinations of SNPs based on phased data from individuals of known origin, we are able to assign individuals of unknown origin to distinct populations, in contrast to using SNPs separately.

DISCUSSION

As genotyping technologies improve, population-genetic datasets increase in number of markers. For example, millions of SNPs have been typed for hundreds of humans (International HapMap 3 Consortium 2010). This development leads to an increase in marker density and substantial levels of LD between many markers. In this study, we focus on how to use dense sets of SNPs for assigning individuals of unknown origin to candidate populations. The idea is to incorporate information from recombination events through combining SNPs into haplotype-loci. We describe a new statistic, the Gain of Informativeness for Assignment from haplotype data, as a decision criterion for combining SNPs into haplotype-loci. GIA compares the informativeness for assignment contained in a haplotype-locus with the sum of the informativeness for assignment contained in each constitutive locus forming the haplotype-locus. If the data consists of genotype data from diploids, a phasing step is needed to infer the phase of the two chromosomes in each individual before GIA can be used to construct a set of haplotype-loci. We show that combining SNPs into haplotype-loci using GIA improves the accuracy of assigning individuals to populations, whereas a strategy of randomly combining SNPs into haplotype-loci leads to less efficient assignment. This result demonstrates that not all haplotypes improve assignment and that combining markers sometimes result in poorer assignment, which may appear surprising since haplotype-loci are multi-allelic and should therefore be more informative about ancestry (compare for example with the use of microsatellites in forensics). However, if we consider the extreme situation where all SNPs are combined into one haplotype-locus, most individuals would have (two) unique haplotype-alleles and the information on ancestry would be nearly zero. There may be an optimum number of SNPs to include in haplotype-loci, but this value will depend on both SNP density and levels of LD, which both vary across the genome. The observed modes for the distribution of number of SNPs in haplotype-loci (figure 8) give an indication of the optimum for the particular cases that we investigate.

We use simulations based on a two-island-model with continuous migration between the populations, and empirical data from the POPRES panel (Nelson et al. 2008), to investigate how different strategies can improve assignment of individuals to populations. Similar to many empirical population studies, the simulated data may contain recent migrants from one population to the other. In our set-up, an individual is considered to be incorrectly assigned when it is not assigned to the population it was sampled from, regardless of whether the individual was a very recent migrant or not. This means that among the individuals deemed incorrectly assigned, there may be a proportion of recent migrants who are justifiably assigned to the population of their recent ancestry (which is not the population they were sampled from). We may therefore expect a small fraction of incorrectly assigned individuals regardless of the as-
assignment approach, but this phenomenon will have little effect on our simulation study. Indeed, for a migration rate $m = 0.01$ and a sample size of 200, we expect 2 individuals being first generation migrants in the sample, with a variance of 2, but this number is too small to explain the high number of incorrectly assigned individuals using, for example, the entire set of SNPs or the pruned set of SNPs (figures 5, 6 and 7).

GIA is well adapted for assignment problems where individuals or segments of genomes are assigned to a population among candidate populations for which we have estimates of allele frequencies for the SNPs and for the haplotype-loci. In particular, a recursive greedy algorithm was found to improve assignment substantially. Interestingly, assignment based the same greedy algorithm, but using $F_{ST}$ (the difference between haplotype-based $F_{ST}$ and single marker-based $F_{ST}$) instead of GIA to determine which markers to combine, also performs much better than assignment based on single SNPs (figure 5). This observation suggests that it is the guided combination of SNPs into haplotypes that leads to the improved assignment and not a particular property of GIA, although GIA is a useful tool for determining which SNPs to combine.

For population structure problems, GIA cannot be used directly because it requires some knowledge about the allele frequencies within the populations, but it could potentially be integrated into MCMC algorithms for estimating population structure, where the algorithm involve a step of partitioning individuals, such as in BAPS (Corander et al. 2003, 2004), TESS (Chen et al. 2007; Durand et al. 2009) or STRUCTURE (Pritchard et al. 2000; Falush et al. 2003). Briefly, for a particular proposed partition, allele frequencies can be estimated from the partitioned sample, and GIA can be computed and used to improve the inference of population structure.

We have demonstrated that haplotypes contain additional information about population structure and that using haplotypes instead of single SNPs can improve assignment of individuals to populations. The GIA statistic determines when it is possible to improve the assignment of individuals to populations by combining markers into haplotypes and it can be used as a tool for population structure inference methods to capitalize on dense sets of genetic markers.

We would like to thank M. Blum, P. Sjödin, C. Schlebusch and two anonymous reviewers for helpful discussions and comments on the manuscript, N. Dufoire-Frebourg for technical comments. The POPRES data were obtained from dbGaP (accession number phs000145.v1.p1). Financial support was provided by the Swedish Research Council and the Swedish Research Council Formas.

REFERENCES


1787–1802.  
Weir, B. S., 1996 Genetic Data Analysis II. Sinauer, Sunderland, MA.  

APPENDIX

We re-write equation 2. Denote the frequency of allele $u$ at locus A in population $i$ by $a_{u}^{i}$, the frequency of allele $v$ at locus B in population $i$ by $b_{v}^{i}$ and the frequency of allele $uv$ of the haplotype-locus, formed by allele $u$ at locus A and allele $v$ at locus B in
population \(i\) by \(x_{uv}^{(i)}\),

\[
GIA = \sum_{u=1}^{U} \sum_{v=1}^{V} \left( -\bar{x}_{uv} \log \bar{x}_{uv} + \sum_{i=1}^{K} \frac{x_{uv}^{(i)}}{K} \log x_{uv}^{(i)} \right) \\
- \sum_{u=1}^{U} \left( -\bar{a}_u \log \bar{a}_u + \sum_{i=1}^{K} \frac{a_u^{(i)}}{K} \log a_u^{(i)} \right) \\
- \sum_{v=1}^{V} \left( -\bar{b}_v \log \bar{b}_v + \sum_{i=1}^{K} \frac{b_v^{(i)}}{K} \log b_v^{(i)} \right), \tag{3}
\]

with \(U\) and \(V\) denoting the number of alleles at locus A and locus B respectively, and using the convention of \(0 \log 0 = 0\).

**Theorem.** Let \(A\) and \(B\) be two bi-allelic loci and \(H\) be their associated haplotype-locus. Consider \(K\) randomly mating populations. For population \(i\), let \(a_u^{(i)}\) and \(b_v^{(i)}\) be the frequencies of the minor allele at locus A and locus B respectively. Then, for all the frequency distributions of the alleles,

\[
\forall i \in 1 \ldots K, \ D_i = 0 \Rightarrow GIA = IA(H) - IA(A) - IA(B) \leq 0
\]

with equality if and only if \(\sum_{i=1}^{K} \sum_{k=1}^{i} (a_u^{(i)} - a_{1}^{(k)})(b_v^{(i)} - b_{1}^{(k)}) = 0\).

**Proof of theorem:** Equation 3 with two bi-allelic loci \((U = 2 \text{ and } V = 2)\) gives

\[
GIA = -\sum_{u=1}^{2} \sum_{v=1}^{2} \bar{x}_{uv} \log \bar{x}_{uv} + \sum_{u=1}^{2} \bar{a}_u \log \bar{a}_u + \sum_{v=1}^{2} \bar{b}_v \log \bar{b}_v \\
+ \frac{1}{K} \sum_{i=1}^{K} \left( \sum_{u=1}^{2} \sum_{v=1}^{2} \frac{x_{uv}^{(i)}}{K} \log x_{uv}^{(i)} - \sum_{u=1}^{2} \frac{a_u^{(i)}}{K} \log a_u^{(i)} - \sum_{v=1}^{2} \frac{b_v^{(i)}}{K} \log b_v^{(i)} \right).
\]

Using the fact that \(a_u^{(i)} = x_{u1}^{(i)} + x_{u2}^{(i)}\) and \(b_v^{(i)} = x_{1v}^{(i)} + x_{2v}^{(i)}\), we obtain

\[
GIA = -\sum_{u=1}^{2} \sum_{v=1}^{2} \bar{x}_{uv} \log \bar{x}_{uv} + \sum_{u=1}^{2} \bar{a}_u \log \bar{a}_u + \sum_{v=1}^{2} \bar{b}_v \log \bar{b}_v \\
+ \frac{1}{K} \sum_{i=1}^{K} \left( \sum_{u=1}^{2} \sum_{v=1}^{2} \frac{x_{uv}^{(i)}}{a_u^{(i)} b_v^{(i)}} \log \frac{x_{uv}^{(i)}}{a_u^{(i)} b_v^{(i)}} \right).
\]

Since all the \(D_i\) are equal to 0, \(x_{uv}^{(i)} = a_u^{(i)} b_v^{(i)}\) for all populations, and \(\log \frac{x_{uv}^{(i)}}{a_u^{(i)} b_v^{(i)}}\), the third term disappears. We define \(\alpha = \bar{x}_{11}, \beta = \bar{x}_{12}, \gamma = \bar{x}_{21} \text{ and } \delta = \bar{x}_{22}\). The \(\alpha, \beta, \gamma\) and \(\delta\) variables are not independent since they sum to 1. Thus, \(GIA\) can be written as a function \(f\) of \(\alpha, \beta\) and \(\gamma\):

\[
GIA = f(\alpha, \beta, \gamma) \\
= (\alpha + \beta) \log(\alpha + \beta) + (\gamma + \delta) \log(\gamma + \delta) + (\alpha + \gamma) \log(\alpha + \gamma) + (\beta + \delta) \log(\beta + \delta) \\
- \alpha \log \alpha - \beta \log \beta - \gamma \log \gamma - \delta \log \delta,
\]

with \(\delta = 1 - \alpha - \beta - \gamma\). The function \(f\) is two-fold differentiable on the open space \(\mathcal{S} = \{\alpha > 0, \beta > 0, \gamma > 0 | \alpha + \beta + \gamma < 1\}\) and we look for the set of points where the gradient of \(f\) is equal to zero, in other words, we are looking for the critical points of \(f\). The first partial derivatives of \(f\) are:

\[
\frac{\partial f}{\partial \alpha}(\alpha, \beta, \gamma) = \log \frac{(\alpha + \beta)(\alpha + \gamma)\delta}{(\delta + \beta)(\delta + \gamma)\alpha} \\
\frac{\partial f}{\partial \beta}(\alpha, \beta, \gamma) = \log \frac{(\alpha + \beta)\delta}{(\delta + \gamma)\beta} \\
\frac{\partial f}{\partial \gamma}(\alpha, \beta, \gamma) = \log \frac{(\alpha + \gamma)\delta}{(\delta + \beta)\gamma}
\]
The first partial derivatives of $f$ are all equal to zero if and only if $\alpha\delta = \beta\gamma$. The nature of the critical points can be investigated by looking at the Hessian matrix $\mathcal{H}$. We can show that for $\alpha\delta = \beta\gamma$, $\mathcal{H}$ can be written as:

$$\mathcal{H} = -\frac{1}{\alpha\delta}X^TX,$$

with $X$ the row vector $(\alpha - \delta, \alpha + \gamma, \alpha + \beta)$ and $X^T$ its transposed vector. $\mathcal{H}$ is thus negative and the critical points defined by $\alpha\delta = \beta\gamma$ are maxima of $f$. Since the equation $\alpha\delta = \beta\gamma$ defines a continuous surface in the open space $S$, defining all values of $S$ on which $f$ reaches a maximum, the value of $f$ on this surface is constant.

$$f(\alpha, \beta, \gamma) = \log \left( \frac{(\alpha + \beta)^{\alpha + \gamma}(\beta + \delta)^{\beta + \delta}(\gamma + \delta)^{\gamma + \delta}}{\alpha\beta\gamma\delta} \right)$$

Using the equality $\alpha\delta = \beta\gamma$, we have

$$(\alpha + \beta)(\alpha + \gamma) = \alpha^2 + (\beta + \gamma)\alpha + \beta\gamma = \alpha(\alpha + \beta + \gamma) + \alpha\delta$$

$$= \alpha$$

Similar computations can be done for the three remaining factors and we find that the maximum value for $f$ on $S$ is therefore 0. This maximum is global on $S$ and since $f$ is extendable by continuity on the border of $S$, it is also a maximum on the closed space $\overline{S}$. Therefore, for all the values of the haplotype-allele frequencies, GIA is less than or equal to zero. Equality is obtained when $\pi_{11}\pi_{22} = \pi_{12}\pi_{21}$.

$$\pi_{11}\pi_{22} = \pi_{12}\pi_{21} \iff \frac{1}{K^2} \left( \sum_{i=1}^{K} a_{1}^{(i)}b_{1}^{(i)} \right) \left( \sum_{k=1}^{K} (1 - a_{1}^{(k)})(1 - b_{1}^{(k)}) \right) = \frac{1}{K^2} \left( \sum_{k=1}^{K} (1 - a_{1}^{(k)})b_{1}^{(k)} \right) \left( \sum_{i=1}^{K} a_{1}^{(i)}(1 - b_{1}^{(i)}) \right)$$

$$\iff \sum_{i=1}^{K} \sum_{k=1}^{K} \left( a_{1}^{(i)}b_{1}^{(i)}(1 - a_{1}^{(k)})(1 - b_{1}^{(k)}) - a_{1}^{(i)}(1 - b_{1}^{(i)})(1 - a_{1}^{(k)})b_{1}^{(k)} \right) = 0$$

$$\iff \sum_{i=1}^{K} \sum_{k=1}^{K} \left( a_{1}^{(i)}(1 - a_{1}^{(k)}) \left( b_{1}^{(i)}(1 - b_{1}^{(k)}) - (1 - b_{1}^{(i)})b_{1}^{(k)} \right) \right) = 0$$

$$\iff \sum_{i=1}^{K} \sum_{k=1}^{K} \left[ a_{1}^{(i)}(1 - a_{1}^{(k)}) \left( b_{1}^{(i)} - a_{1}^{(k)} \right) \right] = 0$$

$$\iff \sum_{i=1}^{K} \sum_{k=1}^{i} \left[ a_{1}^{(i)}(1 - a_{1}^{(k)})b_{1}^{(k)} - b_{1}^{(i)} \right] + \sum_{i=1}^{K} \sum_{k=1}^{K} \left[ a_{1}^{(i)}(1 - a_{1}^{(k)})b_{1}^{(i)} - b_{1}^{(k)} \right] = 0$$

$$\iff \sum_{i=1}^{K} \sum_{k=1}^{i} \left[ a_{1}^{(i)}(1 - a_{1}^{(k)})b_{1}^{(k)} - b_{1}^{(i)} \right] + \sum_{k=1}^{K} \sum_{i=1}^{K} \left[ a_{1}^{(i)}(1 - a_{1}^{(k)})b_{1}^{(i)} - b_{1}^{(k)} \right] = 0$$

At line 5, we add terms for $k = i$ but all those terms are equal to zero. This achieves the proof of theorem. ■