Statistical Applications in Genetics and Molecular Biology

Volume 8, Issue 1

2009

Article 37

Identifying Individuals in a Complex Mixture of DNA with Unknown Ancestry

Joshua Sampson*

Hongyu Zhao[†]

*Yale University, joshua.sampson@yale.edu

[†]Yale University, hongyu.zhao@yale.edu

Identifying Individuals in a Complex Mixture of DNA with Unknown Ancestry*

Joshua Sampson and Hongyu Zhao

Abstract

A new test was recently developed that could use a high-density set of single nucleotide polymorphisms (SNPs) to determine whether a specific individual contributed to a mixture of DNA. The test statistic compared the genotype for the individual to the allele frequencies in the mixture and to the allele frequencies in a reference group. This test requires the ancestries of the reference group to be nearly identical to those of the contributors to the mixture. Here, we first quantify the bias, the increase in type I and type II error, when the ancestries are not well matched. Then, we show that the test can also be biased if the number of subjects in the two groups differ or if the platforms used to measure SNP intensities differ. We then introduce a new test statistic and a test that only requires the ancestries of the reference group to be similar to the individual of interest, and show that this test is not only robust to the number of subjects and platform, but also has increased power of detection. The two tests are compared on both HapMap and simulated data.

KEYWORDS: DNA mixture, genomics, forensics, SNP, identify

*We would like to thank Dr. David Craig for providing us with data from his original article and to the anonymous reviewers for their helpful comments. This work was supported by NIH GM59507.

1 Introduction

Given a mixture of DNA samples from numerous individuals, it is often desirable to determine whether a specific individual contributes DNA to that mixture. Using forensics as an example, this mixture can be a specimen from a crime scene, and the goal can be determining whether a suspect's DNA is included in that specimen. Many methods have been proposed to identify the presence of an individual within a mixture. Most of them focus on cases where only a few people contribute to the mixture. These methods usually compare short tandem repeats (STR) in the mixture to those in the individual (Fung and Hu, 2002; Balding, 2003; Foreman et al., 2003). When only interested in males, the comparison can be limited to STRs on the Y chromosome (Jobling and Gill, 2004). In cases where the DNA has degraded, a second approach, comparing Mitochondrial DNA (mtDNA), specifically the hypervariable region, between mixture and individual can be a better alternative (Stoneking et al., 1991). The limitations of each method were discussed in Homer et al. (2008), but in brief, methods based on STRs require the DNA to be in good condition and methods based on mtDNA, even when augmented by informative SNPs, can have limited discriminatory power (Homer et al., 2008).

In a ground-breaking paper, Homer *et. al.* propose a new method using a genome-wide set of SNPs (Homer et al., 2008) to identify an individual in a mixture. Their method compares $\vec{Y}_M = \{Y_{M1}, \dots, Y_{MN}\}$, where Y_{Mj} is the proportion or frequency of the "A" allele in the mixture at SNP j , with $\vec{Y}_0 = \{Y_{01}, \dots, Y_{0N}\}$, where Y_{0j} is the proportion of the "A" allele in the individual. Clearly, the possible values of Y_{0j} are 0, 0.5, and 1. Note that we have assumed there are N SNPs and have labeled the two possible alleles at each SNP as A and B. Not only is their new method based on a more robust technique, it has higher resolution, able to identify an individual in a mixture containing thousands of subjects. This method also highlighted the potential for a new problem, important enough to be discussed in Science (Couzin, 2008) and require an immediate statement by the NIH. Their results proved that the practice of publicly releasing group-level results from case/control GWAS studies may jeopardize participants' anonymity.

The test statistic, T , proposed by Homer *et. al.*, compares the similarity between $\vec{Y}_0$ and $\vec{Y}_M$ to the similarity between $\vec{Y}_0$ and $\vec{Y}_R$, where $\vec{Y}_R$ are the allele frequencies in a reference mixture. If T is large enough, the corresponding test will reject the null hypothesis that the individual of interest does not contribute DNA to the mixture. For this test to perform well, the subjects for the reference group must be carefully chosen so that their ancestral composition matches that of the subjects contributing to the mixture (e.g. if Caucasians contribute to the mixture, then the reference group must also be Caucasian), or only a select subgroup of ancestry independent SNPs can be used (Kidd et al., 2006). The need for similar reference

groups is clear. Assume we fail to select a similar reference group, and the individual is more similar, in terms of ancestry, to those subjects contributing to the mixture. Then, even if the subject's DNA is absent from the mixture, $\vec{Y}_0$ will be more similar to $\vec{Y}_M$, T will be large, and the test will likely result in a false positive. Similarly, if the individual and the reference group are more similar, the test can lead to a false negative. The obvious problem is that the identities, and therefore the ancestries, of the individuals in the mixture are unknown, and this required matching can be very difficult.

The first goal of this paper is to quantify the magnitude of the bias, type I error, and type II error that can occur if the ancestries of the two groups are poorly matched. Then we identify two other possible sources of bias, the type of platform (e.g. Illumina, Affymetrix) and the number of subjects in the reference group. We use HapMap data to further quantify the extent of the bias in real samples (The International HapMap Consortium, 2003). After demonstrating the severity of the potential problems caused by using T , we propose a new statistic that only requires selecting a reference group with an ancestry that matches the ancestry of the individual of interest. This is a far simpler task, as the individual's ancestry is usually known. Other benefits of the new statistic are that it is unbiased and has a known null distribution. We then compare the performance of the two statistics using simulated data. The paper is therefore ordered as follows. Section II starts by introducing notation and then continues with a discussion of the properties of both statistics and their associated tests. Section III demonstrates the performance of the statistics and their associated tests using both HapMap and simulated data. Finally, section IV contains our brief concluding remarks.

2 Methods

2.1 Notation

Let the individual of interest be indexed by 0, and let the n_R subjects in the reference population and the n_M subjects in the mixture be indexed by $\{1, \dots, n_R + n_M\}$. For subject i , $i \in \{0, 1, \dots, n_R + n_M\}$, let $R_i = 1$ ($M_i = 1$) if subject i is in the reference group (mixture), $R_i = 0$ ($M_i = 0$) otherwise. Unless explicitly stated otherwise, order the subjects so the first n_M belong to the mixture (i.e. $M_i = 1$ for $1 \leq i \leq n_M$). Note that we use the shorthand that "subject i is in the mixture" or "subject i is in the mixture group" if subject i contributes DNA to the mixture. Let e_i be the ancestry, or ethnicity, of subject i . We start by assuming homogenous groups and therefore $e_i \in \{e_0, e_M, e_R\}$, with $e_i = e_R$ if $R_i = 1$, and $e_i = e_M$ if $M_i = 1$.

Let there be N SNPs in the study. Let Q_{ij1} and Q_{ij2} indicate whether the minor allele (relative to ethnicity e_0) is on chromosome 1 and 2, respectively, for sub-

ject i at SNP j . Let $Y_{ij} = 0.5(Q_{ij1} + Q_{ij2})$ be the genotype for subject i at SNP j ; $Y_{ij} \in \{0, 0.5, 1\}$. We define the population allele frequencies by $p_{Mj} \equiv P(Q_{ijk} = 1 | M_i = 1, i > 0)$, $p_{Rj} \equiv P(Q_{ijk} = 1 | R_i = 1, i > 0)$, and $p_{0j} \equiv P(Q_{0jk} = 1)$ where $k \in \{1, 2\}$. For calculating genotype probabilities, we assume Hardy-Weinberg Equilibrium.

Most genotyping platforms directly measure fluorescent intensity, which should be proportional to allele frequency. We shall label the intensity measures for allele A and allele B at SNP j from the mixture group as I_{AMj} and I_{BMj} . These intensities are measurements from a “pooled sample”. We label the ratio of intensities by $\gamma_{Mj} = I_{AMj} / (I_{AMj} + I_{BMj})$. If intensity measurements are also available for individuals, then we can calculate a constant k_{Mj} using known formulas (Pearson et al., 2007; Macgregor et al., 2008) and use a better proxy for the A allele frequency, $\gamma_{Mj} = I_{AMj} / (I_{AMj} + k_{Mj} I_{BMj})$. Similarly, define $\{I_{ARj}, I_{BRj}, \gamma_{Rj}, k_{Rj}\}$ for the reference group. Therefore, the distance measure at each SNP from the Homer *et. al.* paper can be defined as

$$(1) \quad D_{L_1}(j) = |Y_{0j} - \gamma_{Rj}| - |Y_{0j} - \gamma_{Mj}|$$

Their test statistic can then be defined as

$$(2) \quad T_{L_1} \equiv \frac{\sqrt{N}(D_{L_1} - \mu_0)}{\sqrt{N^{-1} \sum_j (D_{L_1}(j) - D_{L_1})^2}}$$

where $D_{L_1} \equiv \frac{1}{N} \sum_j D_{L_1}(j)$ and $\mu_0 = 0$. In this manuscript, we will also discuss

$$(3) \quad D_{L_2}(j) = (Y_{0j} - \gamma_{Rj})^2 - (Y_{0j} - \gamma_{Mj})^2$$

and the corresponding values D_{L_2} and T_{L_2} . The “ L_1 ” and “ L_2 ” subscripts emphasize the distance measure used for the statistic. We discuss the case when intensities perfectly reflect underlying allele frequencies and define $X(j) \equiv |Y_{0j} - \hat{p}_{Rj}| - |Y_{0j} - \hat{p}_{Mj}|$, where $\hat{p}_{Mj} \equiv n_M^{-1} \sum_{i:M_i=1} Y_{ij}$ and $\hat{p}_{Rj} \equiv n_R^{-1} \sum_{i:R_i=1} Y_{ij}$.

In general, we assume that γ_{Mj} and γ_{Rj} can be described by the following statistical models:

$$(4) \quad \begin{aligned} \gamma_{Mj} &= \frac{\sum_{i:M_i=1} Y_{ij}}{n_M} + \beta_{Mj} + \varepsilon_{Mj} = \hat{p}_{Mj} + \beta_{Mj} + \varepsilon_{Mj} \\ \gamma_{Rj} &= \frac{\sum_{i:R_i=1} Y_{ij}}{n_R} + \beta_{Rj} + \varepsilon_{Rj} = \hat{p}_{Rj} + \beta_{Rj} + \varepsilon_{Rj} \end{aligned}$$

where $\varepsilon_{Mj} \sim N(0, \sigma_{Mj}^2)$, $\varepsilon_{Rj} \sim N(0, \sigma_{Rj}^2)$, and $\{\beta_{Mj}, \beta_{Rj}\}$ are SNP/platform specific biases. Therefore, the error is now linearly associated with the ratio of intensities, instead of the intensities themselves. As studies tend to exclude SNPs with rare

minor alleles, we assume that $0 \leq \gamma_{Mj}, \gamma_{Rj} \leq 1$ will generally hold. This description of γ_{Mj} and γ_{Rj} as normal variables is a simplification, but we use it only to show the existence of and then approximate expected biases. Note that the properties of our suggested test statistic do not depend on the validity of equation 4. We define $\vec{\beta}_M \equiv \{\beta_{M1}, \dots, \beta_{MN}\}$, $\vec{\beta}_R \equiv \{\beta_{R1}, \dots, \beta_{RN}\}$, $\vec{\sigma}_M^2 \equiv \{\sigma_{M1}^2, \dots, \sigma_{MN}^2\}$, and $\vec{\sigma}_R^2 \equiv \{\sigma_{R1}^2, \dots, \sigma_{RN}^2\}$. Finally, we use the abbreviation Θ to represent the set of all parameters, $\Theta = \{n_M, n_R, \vec{\beta}_M, \vec{\beta}_R, \vec{\sigma}_R^2, \vec{\sigma}_M^2, e_0, e_M, e_R\}$.

	Ethnicities			
	$M_0 = 1$	$e_0 = e_M$	$e_0 = e_R$	$e_R = e_M$
1 (H_0)		Yes	Yes	Yes
2 (H_1)	Yes	Yes	Yes	Yes
3 (H_{fp})		Yes		
4			Yes	
5				Yes
6				
7	Yes	Yes		

Table 1: There are seven scenarios depending on whether the individual of interest is in the mixture (column 2) and depending on whether the ethnicity of the individual of interest (e_0), the ethnicity of the mixture group (e_M), and the ethnicity of the reference group (e_R) are the same (columns 3,4,5). ‘Yes’ indicates the statement at the top of the column is true and a blank indicates the statement is false. For example, in scenario 1, which we label as “ H_0 ”, the individual of interest is not in the mixture (the statement $M_0 = 1$ is false), the ethnicities of the individual and mixture are the same ($e_0 = e_M$), the ethnicities of the individual and reference are the same ($e_0 = e_R$), and the ethnicities of the mixture and reference are the same ($e_R = e_M$). In this case, all individuals share a common ethnicity. Scenarios 1 and 2 were studied by Homer et al. (2008). Here, we also study scenario 3, which we label as H_{fp} because it is likely to lead to a false positive result.

2.2 D_{L_1} : The Original Test

The definition of any test requires a statement of the null hypothesis and the list of outcomes that would lead to the rejection of that null. Therefore, the desired null hypothesis is that the individual of interest is not in the mixture, $M_0 = 0$, and the original test is to reject this null when T_{L_1} , with $\mu_0 = 0$, is large. The next goal is

to calculate the appropriate threshold, type I error rate, and power. Unfortunately, T_{L_1} is not a pivotal statistic, in that its distribution still depends on the parameter set Θ . Stating $M_0 = 0$ does not lead to a single distribution of T_{L_1} . Therefore, for the original paper to discuss the type I error rate and power, there needed to be three additional assumptions

- I) Identical Ethnicities: $e_0 = e_M = e_R$
- II) Identical Sample Sizes: $n_M = n_R$
- III) Identical Platforms: $\beta_{Mj} = \beta_{Rj}$ and $\sigma_{Mj}^2 = \sigma_{Rj}^2 \forall j$

With these added assumptions, T_{L_1} no longer depends on Θ and they could choose a threshold t_α so that the type I error of the test would be

$$(5) \quad P(T_{L_1} > t_\alpha | M_0 = 0, e_0 = e_M = e_R, n_M = n_R, \vec{\beta}_M = \vec{\beta}_R, \vec{\sigma}_M^2 = \vec{\sigma}_R^2) = \alpha$$

The power was then calculated as

$$(6) \quad P(T_{L_1} > t_\alpha | M_0 = 1, e_0 = e_M = e_R, n_M = n_R, \vec{\beta}_M = \vec{\beta}_R, \vec{\sigma}_M^2 = \vec{\sigma}_R^2) = \beta_{POW}$$

Unfortunately, if one or more of these three assumptions are false, the type I error rate and power can differ from α and β_{POW} . Here, we consider a test to be biased for a given Θ if $P(T_{L_1} > t_\alpha | M_0 = 0, \Theta) \neq \alpha$ or $P(T_{L_1} > t_\alpha | M_0 = 1, \Theta) \neq \beta_{POW}$. In the following sections, we show the magnitude of the bias when each of the assumptions is violated.

2.2.1 D_{L_1} : Bias from Ancestry

In this section, we show that unless the ancestries of the reference and mixture groups are extremely well matched, the false positive rate can easily be near 1, as opposed to the predicted α . Ignoring the sample size and platform parameters, there are seven possible scenarios (Table 1). The original paper thoroughly discussed scenarios H_0 and H_1 with assumptions II and III holding, but only warned of potential bias in the other cases (i.e $P(T_{L_1} > t_\alpha | M_0 = 0, e_M, e_R, e_0) \neq \alpha$ and $P(T_{L_1} > t_\alpha | M_0 = 1, e_M, e_R, e_0) \neq \beta_{POW}$ for all combinations of e_0 , e_M , and e_R). Here, we want to estimate the magnitude of the type I error for the scenario H_{fp} where the result is likely to be a false positive (Table 1). Under H_{fp} , the individual is not in the mixture, but only the ancestries of the individual of interest and the mixture are the same ($e_0 = e_M \neq e_R$).

Estimating the true type I error rate requires understanding the distribution of T_{L_1} , or equivalently, $D_{L_1}(j)$, for the three scenarios, H_0 , H_1 , and H_{fp} . As the allele frequencies at many of the SNPs will likely be the same for all populations,

$p_{0j} = p_{Rj} = p_{Mj}$, the type I error rate will depend on s , defined to be the proportion of all SNPs that are ancestry independent, and the differences between p_{Rj} and p_{Mj} for the other $(1-s)N$ SNPs. To start, we calculate the asymptotic distribution of $D_{L_1}(j)$ (see appendix 5.1), the fundamental component of T_{L_1} , in all three scenarios. When n is large, $n \equiv n_M = n_R$, and $\beta_{Rj} = \beta_{Mj} \forall j$,

$$(7) \quad \begin{aligned} H_0 : D_{L_1}(j) &\sim N(\mu_0(j), \sigma^2(j)) \\ H_1 : D_{L_1}(j) &\sim N(\mu_1(j), \sigma^2(j)) \\ H_{fp} : D_{L_1}(j) &\sim N(\mu_{fp}(j), \sigma_{fp}^2(j)) \end{aligned}$$

where

$$(8) \quad \begin{aligned} \mu_0(j) &= 0 \\ \mu_1(j) &= \frac{1}{n} 2p_{Mj}(1-p_{Mj})^2 \\ \mu_{fp}(j) &= (p_{Mj} - p_{Rj})(1 - 2(1-p_{Mj})^2) \text{ if } p_{Rj} < 0.5 \\ &\quad - p_{Mj}(2p_{Mj}^2 - 6p_{Mj} + 3) + p_{Rj}(1 - 2p_{Mj}^2) \text{ if } p_{Rj} > 0.5 \\ \sigma^2(j) &= \frac{p_{Mj}(1-p_{Mj})}{n} + \sigma_{Mj}^2 + \sigma_{Rj}^2 \\ \sigma_{fp}^2(j) &= \frac{1}{2n} p_{Mj}(1-p_{Mj}) + \frac{1}{2n} p_{Rj}(1-p_{Rj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2 + \\ &\quad (p_{Mj} - p_{Rj})^2 4p_{Mj}(2-p_{Mj})(1-p_{Mj})^2 \end{aligned}$$

Figure 1 shows the type I error rate as a function of s assuming that p_{Mj} is uniformly distributed over the interval $(0,0.5)$, $p_{Rj} = p_{Mj}$ at SNPs $1, \dots, sN$, and p_{Rj} is uniformly distributed over the interval $(0,0.5)$ for the remaining $(1-s)N$ SNPs independent of p_{Mj} . This will overestimate the type I error rate for a given s , because, in truth, p_{Rj} will often be relatively close to p_{Mj} . Nevertheless, this simplified example demonstrates the magnitude of the type I error rate. Figure 1 shows that when n_M is large (i.e. $n_M = 1000$), over 99% of all SNPs would need to be ancestry independent to prevent the false positive rate from being near 1. Therefore, the false positive rate can easily equal or exceed the power of the test. Next, we can calculate the value of s needed for the false positive rate to equal the power, which is essentially when $\mu_1 = \bar{\mu}_{fp}$, where $\mu_1 \equiv E[\mu_1(j)]$ and $\bar{\mu}_{fp} \equiv (1-s)E[\mu_{fp}(j)]$. In appendix 5.2, assuming the allele frequencies have the above uniform distributions, we show that $\mu_1 = 0.23/n$ and $\bar{\mu}_{fp} = 0.062(1-s)$. Therefore, if $1-s = 3.7/n$, $\mu_1 = \bar{\mu}_{fp}$. If our mixture contained 1000 subjects, the reference and mixture groups would only need to differ at 0.37% of all SNPs for the false positive rate to exceed the power. Had we allowed the allele frequencies for those $(1-s)N$ SNPs to be ran-

domly distributed over the interval $[0,1]$, even fewer SNPs would need to differ. Here, $\mu_1 = \bar{\mu}_{fp}$ if $1 - s = 1.2/n$, and only 0.12% of all SNPs would need to differ.

Figure 1: The false positive rate can easily exceed power when tests are based on T_{L1} . Using the normal approximations given in equations 7 and 8, we plot the false positive rate when H_{fp} is true, showing its decrease with the proportion, s , of SNPs that are ancestry independent. Under H_1 , the power was set to be 0.8. Calculations assume $N=50,000$ and $n \in \{100, 1000\}$.

2.2.2 D_{L1} : Bias from Sample Size

In this section, we show that unless the number of subjects in the mixture and reference groups are equal, either type I or type II error can be higher than expected. Here, we keep assumptions I and III, identical ancestries and platforms, and examine the test's bias if n_M and n_R are finite and unequal. Before discussing the origin and extent of this bias, we state a useful fact, which is proven in appendix 5.3.

Inequality of Absolute Values:

Let $p < 0.5$, $p_1 \sim N(p, \sigma_1^2)$, $p_2 \sim N(p, \sigma_2^2)$, and $p_1 \perp p_2$.

If $\sigma_1^2 > \sigma_2^2$, then $E[|0.5 - p_1|] > E[|0.5 - p_2|]$

Figure 2: $E[D_{L_1}(j)] \neq 0$ even when $p_{Rj} = p_{Mj} \equiv p$. In this example, where we let $n_R = \infty$ ($\hat{p}_{Rj} = p$), we find that $E[D_{L_1}(j)]$ decreases as the number of individuals in the mixture shrinks toward 0. We plot $E[D_{L_1}(j)]$ (y-axis) vs n_M (x-axis) for multiple values of p .

Our goal is to show that if $n_M \neq n_R$, then $E[D_{L_1}(j)] \neq 0$ when $M_0 = 0$, even under otherwise ideal conditions. This is demonstrated empirically by figure 2. A more complete discussion follows. Without loss of generality, assume that we have a large reference sample and $n_R > n_M$. Then $\text{var}(\hat{p}_{Mj}) > \text{var}(\hat{p}_{Rj})$. Because $\hat{p}_{Rj}$ is a more precise estimate of p , the expected L_1 distance between Y_{0j} and $\hat{p}_{Rj}$ tends to be smaller than the expected distance between Y_{0j} and $\hat{p}_{Mj}$, or, equivalently, $E[D_{L_1}(j)] < 0$. This would be a trivial statement if distance were measured in L_2 . For L_1 , we will show that $E[X_j] < 0$, which is essentially equivalent to $E[D_{L_1}(j)] < 0$. Recall, $X(j) \equiv |Y_{0j} - \hat{p}_{Rj}| - |Y_{0j} - \hat{p}_{Mj}|$. To use the *Inequality of Absolute Values*, we start by assuming $\hat{p}_{Rj} \sim N(p, n_R^{-1} \sigma_p^2)$ and $\hat{p}_{Mj} \sim N(p, n_M^{-1} \sigma_p^2)$, and decompose $X(j)$ as

$$(9) \quad X(j) = X_{j1}1(Y_{0j} = 1) + X_{j0.5}1(Y_{0j} = 0.5) + X_{j0}1(Y_{0j} = 0)$$

where

$$(10) \quad \begin{aligned} X_{j1} &= |1 - \hat{p}_{Rj}| - |1 - \hat{p}_{Mj}| \\ X_{j0.5} &= |0.5 - \hat{p}_{Rj}| - |0.5 - \hat{p}_{Mj}| \\ X_{j0} &= |0 - \hat{p}_{Rj}| - |0 - \hat{p}_{Mj}| \end{aligned}$$

It is immediately clear that when the individual of interest is not in the mixture, $E[X_{j1}] = E[\hat{p}_{Rj}] - E[\hat{p}_{Mj}] = 0$ and $E[X_{j0}] = E[\hat{p}_{Rj}] - E[\hat{p}_{Mj}] = 0$. With our assumptions of normality and the above inequality, $E[X_{j0.5}] < 0$. Consequently, $E[X_j] < 0$. For an even more intuitive understanding, let $n_R \approx \infty$ and $\hat{p}_{Rj} = p_{Rj} \equiv 0.5$. Then, clearly $|0.5 - \hat{p}_{Rj}| = 0$, and $X_{j0.5} < 0$.

2.2.3 D_{L_1} : Bias from Platform

In this section, we show that unless the allele intensities for the mixture and reference group are measured on the same platform, the test may be biased, often leading to an increase in type II error. Normally, we need not concern ourselves with platform bias, as we can ensure that both samples will be measured on the same platform. Unfortunately, this is not the case here. For the mixture, we have a sample of DNA and we can choose our preferred platform. For the reference group, we often do not have access to actual samples of DNA, and instead base our estimates of allele frequencies on previously recorded genotypes for a group of individuals. Comparing γ_{Mj} and $n_R^{-1} \sum_{i:R_i=1} Y_{ij}$ is equivalent to comparing allele frequencies measured on two different platforms. Also, even if the reference sample were an actual mixture of DNA, that sample might no longer be available for analysis.

First, if $\beta_M \neq \beta_R$, then $E[D_{L_1}] \neq 0$ when $M_0 = 0$, even under otherwise ideal conditions. Simple algebra shows that if model (4) were true,

$$(11) \quad E[D_{L_1}(j)] \approx (\beta_{Mj} - \beta_{Rj})(1 - 2(1 - p_{Mj})^2)$$

Here, we presume $E[D_{L_1}(j)] > 0$ implies $E[D_{L_1}] > 0$, or that $\sum_{j=1}^N \beta_{Mj} \neq \sum_{j=1}^N \beta_{Rj}$. Fortunately, this bias can be easily removed by substituting $\gamma_{Mj}^* \equiv \gamma_{Mj} - \hat{\beta}_{Mj}$ and $\gamma_{Rj}^* \equiv \gamma_{Rj} - \hat{\beta}_{Rj}$ into the equations for $D_{L_1}(j)$, where $\{\hat{\beta}_{Mj}, \hat{\beta}_{Rj}\}$ is an unbiased approximation of $\{\beta_{Mj}, \beta_{Rj}\}$, or by using the appropriate constants k_{Mj} and k_{Rj} for calculating γ_{Mj} and γ_{Rj} .

However, this does not eliminate the potential bias caused by platform. If $\sigma_{Mj}^2 \neq \sigma_{Rj}^2$, then $E[|Y_{0j} - \gamma_{Rj}| - |Y_{0j} - \gamma_{Mj}|] \neq 0$. The origin of this confounding is identical to that for the $n_M \neq n_R$ case. Here, we would assume that $\hat{p}_{Mj} + \varepsilon_{Mj} \sim N(p, \sigma_1^2)$ and $\hat{p}_{Rj} + \varepsilon_{Rj} \sim N(p, \sigma_2^2)$. Next, note that even if $E[\hat{\beta}_{Mj}] = \beta_{Mj}$, $E[\hat{\beta}_{Rj}] = \beta_{Rj}$, and $\sigma_{Mj}^2 = \sigma_{Rj}^2$, there can still be bias if $\text{var}(\hat{\beta}_{Mj}) \neq \text{var}(\hat{\beta}_{Rj})$, as $E[|Y_{0j} - \gamma_{Rj}^*| - |Y_{0j} - \gamma_{Mj}^*|] \neq 0$. The origin of the confounding is easily identified if we assume that $\hat{p}_{Mj} + \hat{\beta}_{Mj} \sim N(p, \sigma_1^2)$ and $\hat{p}_{Rj} + \hat{\beta}_{Rj} \sim N(p, \sigma_2^2)$. The main importance of these statements is that the reference sample cannot simply be the average of known genotypes. In this case $\sigma_{Rj}^2 = 0$ and $\beta_{Rj} = 0$ (i.e. $\text{var}(\hat{\beta}_{Rj}) = 0$). This would create a considerable test bias.

2.3 D_{L_2} : Switching from L_1 to L_2

Using the L_2 distance (recall $D_{L_2}(j) = (Y_{0j} - \gamma_{Rj})^2 - (Y_{0j} - \gamma_{Mj})^2$) greatly simplifies calculations. Under the null hypothesis H_0 , if we keep assumptions II and III, the mean and variance of $D_{L_2}(j)$ can be easily calculated as (see appendix 5.4):

$$\begin{aligned} \mu_0 &= 0 \\ \sigma^2(j) &= 4p_{Mj}^2\sigma_{Mj}^2 + 4p_{Mj}\sigma_{Mj}^2 + 4\sigma_{Mj}^4 - 8p_{Mj}^2\sigma_{Mj}^2 + \\ &\quad \frac{1}{n}(2p_{Mj}^2 - 4p_{Mj}^3 + 2p_{Mj}^4 - 4p_{Mj}^2\sigma_{Mj}^2 + 4p_{Mj}\sigma_{Mj}^2) + \\ &\quad \frac{1}{n^2}(p_{Mj}^2 - 2p_{Mj}^3 + p_{Mj}^4) + \\ (12) \quad &\quad \frac{1}{n^3}\left(\frac{p_{Mj}}{4} - \frac{7}{4}p_{Mj}^2 + 3p_{Mj}^3 - \frac{3}{2}p_{Mj}^4\right) \end{aligned}$$

With the same assumptions, under the alternative hypothesis, H_1 , the mean of $D_{L_2}(j)$ is

$$(13) \quad \mu_1(j) = \frac{1}{n}p_{Mj}(1 - p_{Mj})$$

and the variance can be approximated by $\sigma^2(j)$. In contrast to D_{L_1} , we have estimated the parameters for the normal approximation of D_{L_2} without assuming large n . Equivalently, the approximation of $\sqrt{N}D_{L_2} \sim N(0, N^{-1} \sum_j \sigma^2(j))$ requires only $D_{L_2}(j_1) \perp D_{L_2}(j_2) \forall j_1 \neq j_2$ and N being large. In addition to the ease of L_2 , we found the statistic T_{L_2} to outperform T_{L_1} . This can be verified by simulation (data not shown) or by comparing the two normal approximations for T_{L_1} and T_{L_2} . In the online supplementary material, we compare the power of the statistics T_{L_1} and T_{L_2} for specific values of n , N , $\vec{p}$, and σ^2 , assuming independent SNPs and $p_{Mj} \sim \text{uniform}[0,0.5]$. The plots suggest a mild improvement using T_{L_2} .

2.4 $D_{L_2}^*$: A New Statistic

Our ideal goal would be to develop a pivotal statistic that, when $M_0 = 0$, is a) independent of the individual of interest's ethnicity; b) independent of the number of individuals and the ethnicity of those individuals in the mixture; and c) independent of the platform chosen to analyze the mixture. Although we could find no such statistic, we can take advantage of being able to easily identify an individuals' ethnicity by genotyping a small set of SNPs ($< 0.01N$). The previous requirement of identifying the ethnic composition of the mixture is a far more difficult task. Therefore, we introduce a new statistic that, given e_0 , will have a $N(0,1)$ distribu-

tion when $M_0 = 0$ regardless of the remaining parameters. The key difference in deriving this statistic is that the individuals in the reference group will be selected to have the same ancestry as the individual of interest, as opposed to the same ancestral composition as the mixture. Note that the small set of SNPs used to identify the individuals' ancestries can be removed from the later analyses without greatly diminishing power. A suggested list of SNPs will be made available by the authors in the near future. Moreover, we found that the suggested statistic will lead to a test with increased power. The details of the statistic follow.

To describe the new statistic, we change the notation slightly as we no longer have Y_{ij} values for subjects in the mixture. Therefore, in addition to having γ_{Mj} and Y_{0j} , $1 \leq j \leq N$, choose n_R subjects for a reference group and let Y_{ij} be the genotype for subject i , $1 \leq i \leq n_R$ at SNP j . Note that i is bounded by n_R instead of $n_R + n_M$. Select the reference subjects so their ancestry is similar to that of the individual of interest (i.e. $e_i = e_0$ if $R_i = 1$).

Step 1:

Create $n_R + 1$ reference samples, $\vec{\gamma}_{R0}, \vec{\gamma}_{R1}, \dots, \vec{\gamma}_{Rn_R}$, where $\vec{\gamma}_{Rk} = \{\gamma_{Rk1}, \dots, \gamma_{RkN}\}$, $k \in \{0, \dots, n_R\}$ and

$$(14) \quad \gamma_{Rkj} = \frac{\sum_{i=0, i \neq k}^{n_R} Y_{ij}}{n_R}$$

Here, $\sigma_R^2 = 0$ in model (4). We have immediately removed a source of variation.

Step 2:

Measure the distance between each individual and the appropriate reference group, $(Y_{ij} - \gamma_{Rij})^2$, and compare the resulting value to the distance between that individual and the mixture

$$(15) \quad D_{iL_2}^*(j) = (Y_{ij} - \gamma_{Rij})^2 - (Y_{ij} - \gamma_{Mj})^2$$

Because the σ_R^2 term is absent from $D_{iL_2}^*(j)$, $\text{var}(D_{iL_2}^*(j)) \leq \text{var}(D_{L_2}(j))$. In fact,

$$(16) \quad \text{var}(D_{iL_2}^*(j)) = \text{var}(D_{L_2}(j)) - 2\sigma_{Rj}^2[p_j(1 - p_j) + \frac{p_j(1 - p_j)}{n} + \sigma_{Rj}^2]$$

All variance and covariance calculations will assume $n \equiv n_R = n_M$ and $p_j \equiv p_{Rj} = p_{0j} = p_{Mj}$.

Step 3:

We then average those differences over all reference subjects,

$$(17) \quad \bar{D}_{L_2}^*(j) = \frac{\sum_{i=1}^{n_R} D_{iL_2}^*(j)}{n_R}$$

to obtain an expected difference between the distance to the reference sample and the distance to the mixture, under the null hypothesis. The covariance between two terms in the sum is (see appendix 5.4)

$$(18) \quad \text{cov}(D_{1L_2}^*(j), D_{2L_2}^*(j)) = 2\sigma_{Mj}^4 + \frac{(2p_j - 2p_j^2)\sigma_{Mj}^2}{n} + \frac{\frac{1}{4}p_j - \frac{3}{4}p_j^2 + 2p_j^3 - \frac{1}{2}p_j^4}{n^3} + \frac{\frac{-1}{8}p_j + \frac{11}{8}p_j^2 - \frac{5}{2}p_j^3 + \frac{11}{4}p_j^4}{n^4}$$

Each term on the right side of equation 18 must be positive, so the covariance must be positive (i.e. $\text{cov}(D_{1L_2}^*(j), D_{2L_2}^*(j)) > 0$).

Step 4:

Compare $D_{0L_2}^*(j)$ to this averaged value

$$(19) \quad D_{L_2}^*(j) = D_{0L_2}^*(j) - \bar{D}_{L_2}^*(j)$$

By noting the exchangeability between any two terms $D_{i_1L_2}^*(j)$ and $D_{i_2L_2}^*(j)$, we can calculate the variance of $D_{L_2}^*(j)$

$$(20) \quad \text{var}(D_{L_2}^*(j)) = [1 + \frac{1}{n}][\text{var}(D_{0L_2}^*(j)) - \text{cov}(D_{1L_2}^*(j), D_{2L_2}^*(j))]$$

The key result is that, except for those values of σ_{Mj}^2 where the power is essentially 1 regardless of the chosen statistic, $\text{var}(D_{L_2}^*(j)) < \text{var}(D_{L_2}(j))$. As the means for the L_2 statistics are the same, the smaller variance suggests that our proposed statistic $T_{L_2}^*$, based on $D_{L_2}^*$, will usually have higher power than the original T_{L_2} . This improvement is demonstrated in our simulations. Moreover, the statistic was designed so $E[D_{L_2}^*(j)|H_0] = 0$ if the ancestries of the reference group and individual of interest are matched correctly, regardless of platform and sample size.

Step 5:

Average over all SNPs to get

$$(21) \quad D_{L_2}^* = \frac{\sum_j D_{L_2}^*(j)}{N}$$

Because of the large number of SNPs, the CLT suggests that $\sqrt{N}D_{L_2}^* \sim N(0, \sigma_D^2)$ under the null hypothesis. When allowing for dependency,

$$(22) \quad \text{var}\left(\frac{\sum_j D_{L_2}^*(j)}{\sqrt{N}}\right) = N^{-1} \sum_j \text{var}(D_{L_2}^*(j)) + N^{-1} \sum_{j_1 \neq j_2} \text{cov}(D_{L_2}^*(j_1), D_{L_2}^*(j_2))$$

Therefore, σ_D^2 can be estimated by

$$(23) \quad \hat{\sigma}_D^2 = \frac{\sum_j (D_{L_2}^*(j) - D_{L_2}^*)^2}{N} + \frac{\sum_{j_1 \neq j_2} (D_{L_2}^*(j_1) - D_{L_2}^*)(D_{L_2}^*(j_2) - D_{L_2}^*)}{N}$$

In practice, we restrict the double sum to $j_1 \neq j_2$ and $|j_1 - j_2| \leq 500$. In the on-line supplementary material, we use the HapMap samples to show $N^{-1} \sum_{j_1 \neq j_2} (D_{L_2}^*(j_1) - D_{L_2}^*)(D_{L_2}^*(j_2) - D_{L_2}^*) \approx 0$. We now define our new statistic,

$$(24) \quad T_{L_2}^* \equiv \frac{\sqrt{N}D_{L_2}^*}{\sqrt{\hat{\sigma}_D^2}}$$

and note that $T_{L_2}^* \sim N(0, 1)$. For purposes of future comparisons, we also define $D_{L_1}^*(j)$, $D_{L_1}^*$, and $T_{L_1}^*$ by replacing the L_2 distance with the L_1 distance in $D_{L_2}^*(j)$, $D_{L_2}^*$, and $T_{L_2}^*$.

Although $T_{L_2}^*$ appears to be one of the better performing statistics, it is not necessarily the most intuitive. We first tried a simpler alternative, $D_{L_2}^s$, but as we discuss here, this statistic proved to have extremely low power. Let

$$(25) \quad D_{L_2}^s = \frac{\sum_{j=1}^N D_{L_2}^s(j)}{N}$$

where

$$(26) \quad D_{L_2}^s(j) = (Y_{0j} - \gamma_{Mj})^2 - \frac{\sum_{i=1}^{n_R} (Y_{ij} - \gamma_{Mj})^2}{n_R}$$

This alternative also has the desirable characteristics that neither platform nor sample size can invalidate the equalities $E[(Y_{0j} - \gamma_{Mj})^2] = E[(Y_{ij} - \gamma_{Mj})^2]$ and $E[D_{L_2}^s(j)] = 0$, and that the ancestry of the reference group need only match that of the individual of interest. The problem with such a simple approach is that the $\text{var}(D_{L_2}^s) \approx n \times \text{var}(D_{L_2})$. To understand the origin of this vast difference between variances, we turn to a simple example, where $Y_{ij} \in \{0, 1\}$ and $P(Y_{ij} = 1) = p_{jM}$. Here, $D_{L_2}(j) \in \{2(\gamma_{Mj} - \gamma_{Rj}) + \gamma_{Rj}^2 - \gamma_{Mj}^2, \gamma_{Rj}^2 - \gamma_{Mj}^2\}$, and the $\text{var}(D_{L_2}(j))$, under H_0 , asymptotically approaches 0. In contrast, $D_{L_2}^s(j) \in \{(1 - \gamma_{Mj})^2 + k(\vec{Y}_R, \gamma_{Mj}), \gamma_{Mj}^2 +$

$k(\vec{Y}_R, \gamma_{Mj})\}$, where $k(\vec{Y}_R, \gamma_{Mj})$ is the appropriate function of the reference group. Without the cancelation of the Y_{0j} terms in the statistic, asymptotically, $\text{var}(D_L^s(j)) \rightarrow p_{jM}(1 - p_{jM})(1 + 4p_{jM}^2)$. Therefore, we designed $D_{L_2}^*$ to not only avoid potential bias, but retain the advantage gained by the cancelation of Y_{0j} s in D_{L_1} . Here we also note that the variance of $D_{L_2}(j)$ or any similar statistic will be minimized when $E[\gamma_{Mj}] = E[\gamma_{Rj}]$.

This same simple example, where $Y_{ij} \in \{0, 1\}$ and $P(Y_{ij} = 1) = p_{jM}$, illustrates why the original statistic, $D_{L_1}(j) = |Y_{0j} - \gamma_{Mj}| - |Y_{0j} - \gamma_{Rj}|$, or any of our adaptations, needs to include the reference term, $|Y_{0j} - \gamma_{Rj}|$. Under the null hypothesis, $\text{var}(|Y_{0j} - \gamma_{Rj}| - |Y_{0j} - \gamma_{Mj}|) = 2p_{jM}(1 - p_{jM})/n \rightarrow 0$. In contrast, without the cancelation of Y_{0j} terms, $\text{var}(|Y_{0j} - \gamma_{Mj}|) \rightarrow p_{jM} - 5p_{jM}^2 + 8p_{jM}^3 - 4p_{jM}^4$. Even if the ethnicities were well matched, the variance of the genome-wide statistic would be proportional to the sum of these variances over the entire genome, not just over the subset of informative SNPs. For choosing the best statistic, the variance will be the deciding value. The additional term $|Y_{0j} - \gamma_{Rj}|$ does not change the expected difference between the test statistic under the null and alternative hypotheses.

2.5 Data and Simulations

The power and type I error for the test based on T_{L_1} has already been thoroughly described when $e_0 = e_R = e_M$. Here, we use simulations to accomplish two goals. First, we show that tests based on $T_{L_2}^*$ are more powerful than tests based on T_{L_1} . Second, we show that if the threshold t_α is chosen so that the type I error rate is α when $e_0 = e_R = e_M$ (equation 5), but, in fact, assumption I is not true, then the true type I error rate can greatly exceed α .

The simulations generated distributions of T_{L_1} , T_{L_2} , $T_{L_1}^*$, and $T_{L_2}^*$ for each of the three scenarios, H_0 , H_1 , and H_{fp} . Recall that the denominator for T_{L_1} was $\sqrt{N^{-1} \sum_j (D_{L_1}(j) - D_{L_1})^2}$, which assumes independence across SNPs. Therefore, in order to provide a fair comparison, we let the genotypes be independently distributed and let the denominator for $T_{L_2}^*$ be $\sqrt{N^{-1} \sum_j (D_{L_2}^*(j) - D_{L_2}^*)^2}$.

First, we simulated 10,000 datasets (and 10,000 values of each test statistic) under H_0 , with $e_0 = e_M = e_R$, $n_M = n_R$, and $\beta_M = \beta_R = 0$. Each dataset is simulated as follows. For each of 5000 SNPs (when $n_M = 100$) or 50,000 SNPs ($n_M = 1000$), we randomly generated p_{0j} , the frequency of allele 'A' for the individual of interest, from a uniform(0.1,0.5) distribution. Then, we set $p_{Mj} = p_{R_0j} = p_{R_Mj} = p_{0j}$, where p_{Mj} , p_{R_0j} , and p_{R_Mj} are the 'A' allele frequencies for subjects in the mixture, reference group R_0 (selected to match the individual of interest), and reference group R_M (selected to match the mixture). Given the allele frequencies and assuming

Hardy-Weinberg Equilibrium, we then generated the genotypes for the individual of interest ($\vec{Y}_0$), the n_m individuals in the mixture ($\mathbf{Y}_M$), and the n_R individuals in each reference group ($\mathbf{Y}_{R_0}, \mathbf{Y}_{R_M}$). Given knowledge of the individuals in the mixture and reference groups, we generated the microarray intensities at each SNP as $\gamma_{Mj} = n_M^{-1} \sum_{i:M_i=1} Y_{ij} + \varepsilon_{Mj}$ and $\gamma_{Rj} = n_R^{-1} \sum_{R_{0i}=1} Y_{ij}$, where $\varepsilon_{Mj} \sim N(0, \sigma_M^2)$. Simulations were performed for a range of σ_M^2 values. The results are a single dataset $\{\vec{Y}_0, \vec{\gamma}_M, \mathbf{Y}_{R_0}, \vec{\gamma}_R\}$ and a set of values $\{T_{L_1}, T_{L_2}, T_{L_1}^*, T_{L_2}^*\}$. After simulating all 10,000 datasets under H_0 , we denote the 99th percentile of the four test statistics by $\hat{t}_{\alpha L_1}, \hat{t}_{\alpha L_2}, \hat{t}_{\alpha L_1}^*,$ and $\hat{t}_{\alpha L_2}^*$.

Next, we simulated 1000 datasets under H_1 . To create the datasets, we repeated the above steps, but let $\gamma_M(j) = n_M^{-1} (Y_{0j} + \sum_{i=2}^{n_M} Y_{ij}) + \varepsilon_{Mj}$. Then the power for the original test was the proportion of these 1000 T_{L_1} values exceeding $\hat{t}_{\alpha L_1}$. A similar estimation of power was performed for $T_{L_2}, T_{L_1}^*,$ and $T_{L_2}^*$. Finally, we simulated datasets under H_{fp} . The type I error rate will depend on how the allele frequencies differ between the two ancestries. In one case, we assumed that a proportion, s , of the SNPs to be ancestry independent, (i.e. $p_{Mj} = p_{R_0j} = p_{R_Mj} = p_{0j}$). For $j > s \times 5000$ or $j > s \times 50,000$, we still enforced $p_{Mj} = p_{R_0j} = p_{0j}$, but generated p_{R_Mj} from a uniform(0.1,0.5) distribution. In the second case, we assumed that the allele frequencies differed at all SNPs and let $p_{R_Mj} = (1 + k_j)^{\pm 1} p_{0j}$, where $\{k_1, \dots, k_N\}$ was a vector of equally spaced points over the range 0 to 0.15, 0.20, and 0.25 when $n_M = 100$ and 0 to 0.03, 0.05, and 0.07 when $n_M = 1000$. The superscript ± 1 indicates that the exponent was randomly generated to be -1 or 1 with equal probability. Here, we can quantify the extent of the difference between the two ethnicities by the measure λ , the variance inflation factor typical in case/control studies. For $n_M = 100$, the three ranges correspond to a λ of 1.16, 1.36, and 1.65, and for $n_M = 1000$ the three ranges correspond to 1.09, 1.32, and 1.67. For each of these cases, we simulated 1000 datasets and defined the type I error rate to be the percentage of these 1000 T_{L_1} (T_{L_2}) values exceeding $\hat{t}_{\alpha L_1}$ ($\hat{t}_{\alpha L_2}$).

To better quantify the type I error rate possible in real experiments, we used the 90 CEU and 45 Japanese individuals from the HapMap samples. For each CEU individual i , we selected 9 unrelated CEU individuals to create a positive mixture, $\gamma_{M+} = 0.1(Y_i + \sum_{k=1}^9 \vec{Y}_k) + \vec{\varepsilon}_i$, or 10 unrelated individuals to create a negative mixture, $\gamma_{M-} = 0.1(\sum_{k=1}^{10} \vec{Y}_k) + \vec{\varepsilon}_i$, where $\vec{\varepsilon}_i \equiv \{\varepsilon_{ij}, \dots, \varepsilon_{iN}\}$, $\varepsilon_{ij} \sim N(0, 0.01^2)$ and $\varepsilon_{ij_1} \perp \varepsilon_{ij_2}$ if $j_1 \neq j_2$. To achieve meaningful levels of power, we chose $N = 1000$. For each CEU individual, 11 reference groups were similarly created where γ_{Rt} included t Japanese individuals and $10 - t$ CEU individuals. We calculated the distribution of T_{L_1} (36000 simulations, 90 subjects $\times$ (400,000 SNPs/1000)=400 sets of SNPs) under H_0, H_1 , and H_{fp} . We chose the threshold $\hat{t}_{\alpha L_1}$ so that $\alpha = 0.05$ if H_0 were true and then calculated the probability of rejection under H_1 (i.e. power) and under H_{fp} (i.e. type I error).

3 Results

3.1 Comparison of Power

Our first goal is to show that tests based on $T_{L_2}^*$ are more powerful than tests based on T_{L_1} . As the results for the two sets of simulations, $n_M = n_R = 1000$ and $n_M = n_R = 100$ were similar, we limit our discussion to the latter. Recall, we chose to empirically estimate the 99th percentile of the test statistics under H_0 . This was a necessity because our reference mixture is an average of known genotypes, $\sigma_M^2 \neq \sigma_R^2$, and as discussed in section 2.2.3, T_{L_1} cannot be expected to follow a $N(0,1)$ distribution. Under H_0 , when $\sigma_M = 0.03, 0.05, 0.07$, and 0.09 , the averages of T_{L_1} over the 10,000 simulations were $-0.76, -1.71, -2.77$, and -3.91 respectively. Similarly, the averages of T_{L_2} were $-1.85, -4.13, -6.46$, and -8.68 for those same values of σ_M . The empirical variances were close to 1. In contrast, $T_{L_2}^*$ followed a $N(0,1)$ distribution, as promised, and this empirical step was superfluous. The empirical means for $T_{L_1}^*$ and $T_{L_2}^*$ were near 0, $-0.015 \leq \bar{T}_{L_1}^* \leq -0.010$ and $-0.024 \leq \bar{T}_{L_2}^* \leq -0.015$, regardless of σ_M . Moreover, not only were the empirical variances near 1, the empirical 99th percentiles were approximately 2.326 ($2.29 \leq \hat{t}_{\alpha L_1}^* \leq 2.32$ and $2.31 \leq \hat{t}_{\alpha L_2}^* \leq 2.32$).

For each value of σ_M , we then calculated the power for the four tests, based on $T_{L_1}, T_{L_2}, T_{L_1}^*$, and $T_{L_2}^*$. The alternative hypothesis was H_1 with the other parameters unchanged. These values are plotted against signal noise or σ_M in Figure 3. Clearly, power is larger when the statistics are based on L_2 and when the reference group is matched to the individual of interest. Note, had we included an error term in the reference mixture (and possibly attained $T_{L_1}, T_{L_2} \sim N(0, 1)$), the benefit of using the new statistic would have been larger.

3.2 Examination of False Positive Rate

Our second goal is to show that when assumption I is violated, the type I error rate can be much larger than α . Again, we presume t_α was selected so equation 5 holds. Simulations showed the quick increase in the false positive rate, $P(T_{L_1} > t_\alpha | H_{fp})$, as the proportion of ancestry-independent SNPs decreased in those tests based on T_{L_1} . Parameter values were chosen so that the ideal test ($e_0 = e_M = e_R$), for both $n_M = 100$ and $n_M = 1000$, would have had power = 0.87 to detect an individual's presence. Clearly, unless s is close to 1, the false positive rate exceeds this estimated power. Because we have finite sample sizes and $\sigma_M^2 \neq \sigma_R^2$, we cannot expect the predicted power and false positive rate to be equal when $1 - s = 3.7/n$ (Appendix 5.2). However, the simulations can confirm that the two are approximately equal when $\mu_1 = \mu_{fp}$ and that value of $1 - s$ which results in equality is proportional to $1/n$. When the power and false positive rate cross in Figure 4 for $n_M = 100$, we

Figure 3: Tests based on $T_{L_2}^*$ have the highest power. Power for tests based on the four statistics, T_{L_1} , T_{L_2} , $T_{L_1}^*$, and $T_{L_2}^*$, are plotted for multiple values of σ_M^2 or noise. Other parameters in the simulation were $n_M = n_R = 100$, $N = 5000$, $s = 1$, and $\beta_M = \beta_R = 0$.

find $\mu_1 = 2.40$, $\mu_{fp} = 2.41$, and $1 - s = 0.088$. Similarly for $n_M = 1000$, we find $\mu_1 = 2.65$, $\mu_{fp} = 2.64$, and $1 - s = 0.0089$. For simulations under the second set of p_{RMj} , we found the false positive rates to be 0.11, 0.30, and 0.61 when $n_M = 100$ and 0.035, 0.15, and 0.51 when $n_M = 1000$.

The false positive rate for the HapMap samples were calculated when the mixture and reference groups each had 10 subjects. Both the individual of interest and the subjects in the mixture were from the CEU population. As the ratio of Japanese:CEU individuals increased in the reference group, the false positive rate increased from the α -level, 0.05, to near 1. When the ratio exceeded 8:2, the false positive rate exceeded the estimated power.

4 Discussion

As the overall popularity of SNP microarray technology increases and the cost of the technology decreases, there will likely be a shift from STRs to genome-wide sets of SNPs as the preferred method for DNA identification. Therefore, databases designed to store genetic identifiers for individuals will likely store genotypes for

Figure 4: The type I error rate for tests based on T_{L_1} can exceed power. Type I error rate increases as the proportion, s , of ancestry independent SNPs decreases. Other parameters in the simulation were $N=5000$ (50,000), $\sigma_M^2 = 0.035^2$ (0.01^2), and $n_M = n_R = 100$ (1000).

sets of SNPs in the future. Coupled with our earlier discussion, there will be three main advantages of using high-density SNPs to determine whether an individual contributes DNA to a mixture: accessibility, higher resolving power and the ability to work with low quality, degraded, DNA.

Here, we have demonstrated that tests based on T_{L_1} can suffer from inflated type I and type II errors if the mixture contains individuals with unknown ancestries. Therefore, we introduced a new test statistic, $T_{L_2}^*$ and an accompanying test that only require matching the ancestry of the reference group to that of the individual of interest. Even if the individual of interest has a mixture of ancestries, there should still be some subjects, with a similar mixture of ancestries, that can be used for comparison. This test is also robust to platform and sample size. We showed that both switching from the L_1 to the L_2 measure and switching to the new type of statistic increased the power to detect the individual of interest. Therefore, $T_{L_2}^*$ is not only more robust than T_{L_1} , it tends to have increased power.

Figure 5: The type I error rate for tests based on T_{L_1} can exceed power in simulations based on HapMap samples. The individual of interest and individuals in the mixture group were from the CEU population. The individuals in the reference group were a mixture from the CEU and Japanese populations. The type I error rate increased with the percentage of individuals in the reference group that were from the Japanese population. Other parameters in the simulation were $N=1000$, $n_M = n_R = 10$, and $\sigma_M^2 = \sigma_R^2 = 0.01^2$

5 Appendix

5.1 Distribution of $D_{L_1}(j)$

To estimate the distribution of $D_{L_1}(j)$, we make the following assumptions, where $g \in \{R, M\}$:

- 1) Model (4) is true.
 - 2) $\beta_{Rj} = \beta_{Mj} = 0$
 - 3) $n \equiv n_M = n_R$
- And restrict the SNPs examined to those satisfying
- 4) $|p_{gj} - 0.5| > \delta$
- for some small δ such that
- 5) $P(|\hat{p}_{gj} - p_{gj}| > \delta/2) \approx 0$.
 - 6) $P(|\varepsilon_{gj}| > \delta/2) \approx 0$.

Assumptions 5 and 6 promise that n is large enough and the magnitude of ε_{gj} is small enough so that the signs inside the absolute values are determined by $Y_{0j} - p_{gj}$. For much of the exposition, we will still leave the β_{Rj} and β_{Mj} terms in the equations, although we assume they are small enough to not affect the sign inside the absolute values, to demonstrate their effects on bias and variance had they not been assumed to equal 0. To start, divide $D_{L_1}(j)$ into three components

$$(27) \quad D_{L_1}(j) \approx D_{j1}1(Y_{0j} = 1) + D_{j0.5}1(Y_{0j} = 0.5) + D_{j0}1(Y_{0j} = 0)$$

where

$$\begin{aligned} D_{j1} &\equiv \sum_{i=(n+1)}^{2n} \frac{1-Y_{ij}}{n} - \frac{1-Y_{1j}}{n} - \sum_{i=2}^n \frac{1-Y_{ij}}{n} - \beta_{Rj} - \varepsilon_{Rj} + \beta_{Mj} + \varepsilon_{Mj} \\ D_{j0.5} &\equiv \sum_{i=(n+1)}^{2n} \frac{0.5-Y_{ij}}{n} - \frac{0.5-Y_{1j}}{n} - \sum_{i=2}^n \frac{0.5-Y_{ij}}{n} - \beta_{Rj} - \varepsilon_{Rj} + \beta_{Mj} + \varepsilon_{Mj} \text{ if } p_{Rj} < 0.5 \\ D_{j0.5} &\equiv \sum_{i=(n+1)}^{2n} \frac{Y_{ij}-0.5}{n} - \frac{0.5-Y_{1j}}{n} - \sum_{i=2}^n \frac{0.5-Y_{ij}}{n} + \beta_{Rj} + \varepsilon_{Rj} + \beta_{Mj} + \varepsilon_{Mj} \text{ if } p_{Rj} > 0.5 \\ (28) D_{j0} &\equiv \sum_{i=(n+1)}^{2n} \frac{Y_{ij}}{n} - \frac{Y_{1j}}{n} - \sum_{i=2}^n \frac{Y_{ij}}{n} + \beta_{Rj} + \varepsilon_{Rj} - \beta_{Mj} - \varepsilon_{Mj} \end{aligned}$$

By definition $p_{Mj} < 0.5$. We next calculate the expected values of each component under the two assumptions, H_1 and H_{fp} . We know that $E[Y_{ij}] = p_{gj}^2 + p_{gj}(1 - p_{gj}) = p_{gj}$, where g indicates group. Therefore, assuming scenario H_1 , where without loss of generality, we let the individual of interest be subject 1 of the mixture,

$$(29) \quad E[D_{j1}|H_1] = \frac{1}{n} - \frac{p_{Mj}}{n} - \beta_{Rj} + \beta_{Mj}$$

$$(30) \quad E[D_{j0.5}|H_1] = \frac{1}{2n} - \frac{p_{Mj}}{n} - \beta_{Rj} + \beta_{Mj}$$

$$(31) \quad E[D_{j0}|H_1] = \frac{p_{Mj}}{n} + \beta_{Rj} - \beta_{Mj}$$

We immediately see that

$$(32) \quad nE[D_{L_1}(j)|H_1] = 2p_{Mj}(1 - p_{Mj})^2 + n(\beta_{Mj} - \beta_{Rj})(1 - 2(1 - p_{Mj})^2)$$

Let us assume that $\beta_{Mj} = \beta_{Rj}$, then we get

$$(33) \quad \sum_j E[D_{L_1}(j)|H_1] = \frac{\sum_j 2p_{Mj}(1 - p_{Mj})^2}{n}$$

Next we calculate the expected value under H_{fp} ,

$$\begin{aligned}
 E[D_{j1}|H_{fp}] &= p_{Mj} - p_{Rj} - \beta_{Rj} + \beta_{Mj} \\
 E[D_{j0.5}|H_{fp}] &= p_{Mj} - p_{Rj} - \beta_{Rj} + \beta_{Mj} \text{ if } p_{Rj} < 0.5 \\
 E[D_{j0.5}|H_{fp}] &= p_{Mj} + p_{Rj} - 1 + \beta_{Rj} + \beta_{Mj} \text{ if } p_{Rj} > 0.5 \\
 (34) \quad E[D_{j0}|H_{fp}] &= p_{Rj} - p_{Mj} + \beta_{Rj} - \beta_{Mj}
 \end{aligned}$$

We immediately see that

$$\begin{aligned}
 E[D_{L1}(j)|H_{fp}, p_{Rj} < 0.5] &= (p_{Mj} - p_{Rj} + \beta_{Mj} - \beta_{Rj})(1 - 2(1 - p_{Mj})^2) \\
 E[D_{L1}(j)|H_{fp}, p_{Rj} > 0.5] &= -p_{Mj}(2p_{Mj}^2 - 6p_{Mj} + 3) + p_{Rj}(1 - 2p_{Mj}^2) + \\
 (35) \quad &\quad \beta_{Mj}(1 - 2(1 - p_{Mj})^2) + \beta_{Rj}(1 - 2p_{Mj}^2)
 \end{aligned}$$

For simplification, let us assume that $\beta_{Mj} = \beta_{Rj}$, then we get

$$\begin{aligned}
 \sum_j E[D_{L1}(j)|H_{fp}] &= \sum_{j:p_{Rj} < 0.5} (p_{Mj} - p_{Rj})(1 - 2(1 - p_{Mj})^2) + \\
 (36) \quad &\quad \sum_{j:p_{Rj} > 0.5} -p_{Mj}(2p_{Mj}^2 - 6p_{Mj} + 3) + p_{Rj}(1 - 2p_{Mj}^2)^2
 \end{aligned}$$

Now, we turn our attention to the variance of $D_{L1}(j)$ and recall that

$$(37) \quad \text{var}(D_{L1}(j)) = E[\text{var}(D_{L1}(j)|Y_{0j})] + \text{var}(E[D_{L1}(j)|Y_{0j}])$$

We calculate the $\text{var}(D_{L1}(j)|Y_{0j})$, assuming all individuals within a group are unrelated,

$$\begin{aligned}
 E[\text{var}(D_{L1}(j)|Y_{0j})] &= \frac{1}{2n} p_{Rj}(1 - p_{Rj}) + \frac{\text{var}(Y_{1j})}{n^2} + \\
 (38) \quad &\quad \frac{n-1}{2n^2} p_{Mj}(1 - p_{Mj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2
 \end{aligned}$$

Under H_0 , where $E[D_{L1}(j)|Y_{0j}] = 0$ and $p_{Rj} = p_{Mj}$, we see that

$$(39) \quad \text{var}(E[D_{L1}(j)|Y_{0j}, H_0]) = [(\beta_{Mj} - \beta_{Rj})^2 4p_{Mj}(2 - p_{Mj})(1 - p_{Mj})^2]$$

For simplification, we again assume that $\beta_{Mj} = \beta_{Rj}$ and therefore

$$(40) \quad \text{var}(D_{L1}(j)|H_0) = \frac{1}{n} p_{Mj}(1 - p_{Mj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2$$

Under H_1 , we find that

$$(41) \quad E[\text{var}(D_{L_1}(j)|Y_{j0}, H_1)] = \frac{2n-1}{2n^2} p_{Mj}(1-p_{Mj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2$$

and for large n , we see that $E[\text{var}(D_{L_1}(j)|Y_{j0}, H_1)] \approx E[\text{var}(D_{L_1}(j)|Y_{j0}, H_0)]$. Similarly, because $E[D_{L_1}(j)|Y_{0j}, H_1] = E[D_{L_1}(j)|Y_{0j}, H_0] + O(\frac{1}{n})$, then $\text{var}(E[D_{L_1}(j)|Y_{0j}, H_0]) = \text{var}(E[D_{L_1}(j)|Y_{0j}, H_1]) (1 + O(\frac{1}{n})) + O(\frac{1}{n^2})$. Therefore, we can assume that $\text{var}(D_{L_1}(j)|H_1) \approx \text{var}(D_{L_1}(j)|H_0)$.

Next, we turn to scenario H_{fp} . Clearly,

$$(42) \quad E[\text{var}(D_{L_1}(j)|Y_{0j}, H_{fp})] = \frac{1}{2n} p_{Mj}(1-p_{Mj}) + \frac{1}{2n} p_{Rj}(1-p_{Rj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2$$

To calculate the $\text{var}(E[D_{L_1}(j)|Y_{0j}, H_{fp}])$, we use the notation $p_{\Delta j} = p_{Mj} - p_{Rj}$ and $\beta_{\Delta j} = \beta_{Mj} - \beta_{Rj}$, then under H_{fp} , if we let $p_{Rj} < 0.5$,

$$(43) \quad \begin{aligned} \text{var}(E[D_{L_1}(j)|Y_{0j}, H_{fp}]) = \\ [(p_{\Delta j} - \beta_{\Delta j})^2 4p_{Mj}(2-p_{Mj})(1-p_{Mj})^2] \end{aligned}$$

Now, if we assume that the minor allele in the mixture is the same as the minor allele in the reference group, or equivalently, that we have at least identified a reference population of moderately similar composition, then we can be satisfied that equation (43) is a satisfactory approximation of $\text{var}(E[D_{L_1}(j)|Y_{0j}, H_{fp}])$. Again, with the simplification that $\beta_{Mj} = \beta_{Rj}$, we find

$$(44) \quad \begin{aligned} \text{var}(D_{L_1}(j)|H_{fp}) = \frac{1}{2n} p_{Mj}(1-p_{Mj}) + \frac{1}{2n} p_{Rj}(1-p_{Rj}) + \sigma_{Rj}^2 + \sigma_{Mj}^2 + \\ [(p_{Mj} - p_{Rj})^2 4p_{Mj}(2-p_{Mj})(1-p_{Mj})^2] \end{aligned}$$

5.2 Estimating when $\mu_1 = \bar{\mu}_{fp}$

Here, we provide a rough approximation of the percentage of SNPs that would need to be ancestry dependent for $E[\mu_1(j)] = (1-s)E[\mu_{fp}(j)]$. We base this approximation on the assumption that p_{Mj} and p_{Rj} are uniform random variables over the intervals $[0,0.5]$ and $[0,1]$ respectively. Start by calculating $E[D_{L_1}(j)|p_{Rj} < 0.5]$ from equation 35, letting $\beta_{Mj} = \beta_{Rj}$ where the expectation is over p_{Rj} and p_{Mj} .

$$\begin{aligned}
 & E[D_{L_1}(j)|p_R < 0.5, H_{fp}] \\
 &= 4 \int_0^{0.5} \int_0^{0.5} (p_{Mj} - p_{Rj})(4p_{Mj} - 1 - 2p_{Mj}^2) dp_{Rj} dp_{Mj} \\
 &= 4 \int_0^{0.5} \int_0^{0.5} (-4p_{Rj}p_{Mj} + 4p_{Mj}^2 + p_{Rj} - p_{Mj} + 2p_{Rj}p_{Mj}^2 - 2p_{Mj}^3) dp_{Rj} dp_{Mj} \\
 (45) \quad &= 0.062
 \end{aligned}$$

Had we assumed $p_{Rj} > 0.5$, $E[D_{L_1}(j)|p_{Rj} > 0.5, H_{fp}] = 0.32$. Combining the two equalities, we find $E[D_{L_1}(j)|H_{fp}] = 0.19$. Obviously, from an evolutionary perspective, p_R and p_M should often be similar in value, but we ignore that here for simplicity. Furthermore, assume that only some proportion, $1 - s$, of those N SNPs differ among ethnicities. Then $E[\mu_{fp}(j)] = 0.19$ and $\mu_{fp} \equiv (1 - s)E[\mu_{fp}(j)] = 0.19(1 - s)$.

Next, from equation 33, we know $nE[D_{L_1}(j)] = 2p_{Mj}(1 - p_{Mj})^2$, and again assuming p_{Mj} is a uniform random variable over the interval $[0, 0.5]$, we find $E[D_{L_1}(j)|H_1] = .22/n$. Therefore, across all SNPs, assuming no linkage, $\mu_1 \equiv E[\mu_1(j)] = .22/n$.

Therefore, in order for $\mu_1 = \bar{\mu}_{fp}$, we would need $.22/n = 0.19(1 - s)$, which shows that false positives can be expected so long as the percentage of ancestry dependent SNPs, $1 - s$, exceeds $1.16/n$. Remember, that the difference between p_R and p_M is likely to be overstated by the above simplifications, and therefore this $1.16/n$ will be low.

5.3 Inequality of Absolute Values (proof)

Theorem: Let $X_1 \sim N(p, \sigma_1^2)$, $X_2 \sim N(p, \sigma_2^2)$, $g(X) = |0.5 - X|$, $X_1 \perp X_2$, $\sigma_1^2 > \sigma_2^2$, and $p < 0.5$. Then

$$(46) \quad E[g(X_1)] > E[g(X_2)]$$

Let $Z_1 = F_1(X_1)$ and $Z_2 = F_2(X_2)$, where F_1 and F_2 are their respective cumulative distribution functions. Then

$$\begin{aligned}
 E[g(X_2)] - E[g(X_1)] &= \int_{-\infty}^{\infty} g(X_2) f_2(X_2) dX_2 - \int_{-\infty}^{\infty} g(X_1) f_1(X_1) dX_1 \\
 &= \int_0^1 g(F_2^{-1}(Z_2)) dZ_2 - \int_0^1 g(F_1^{-1}(Z_1)) dZ_1 \\
 &= \int_0^{0.5} g(F_2^{-1}(Z_2)) + g(F_2^{-1}(1 - Z_2)) dZ_2 - \\
 (47) \quad &\int_0^{0.5} g(F_1^{-1}(Z_1)) + g(F_1^{-1}(1 - Z_1)) dZ_1
 \end{aligned}$$

There are two possibilities. Case A) Assume $F_2^{-1}(1 - Z_2) < 0.5$. Then, by linearity of $g(X)$,

$$(48) \quad \frac{g(F_2^{-1}(Z_2)) + g(F_2^{-1}(1 - Z_2))}{2} = \frac{g(F_2^{-1}(Z_2) + F_2^{-1}(1 - Z_2))}{2} = g(p) = 0.5 - p$$

Clearly, by similar logic,

$$(49) \quad \frac{g(F_1^{-1}(Z_1)) + g(F_1^{-1}(1 - Z_1))}{2} \geq 0.5 - p = \frac{g(F_2^{-1}(Z_2)) + g(F_2^{-1}(1 - Z_2))}{2}$$

Case B: $F_2^{-1}(1 - Z_2) \geq 0.5$. Then, clearly

$$\begin{aligned}
 g(F_1^{-1}(1 - Z_1)) &> g(F_2^{-1}(1 - Z_2)) \\
 (50) \quad g(F_1^{-1}(Z_1)) &> g(F_2^{-1}(Z_2))
 \end{aligned}$$

Therefore, for case B we also see

$$(51) \quad \frac{g(F_1^{-1}(Z_1)) + g(F_1^{-1}(1 - Z_1))}{2} \geq \frac{g(F_2^{-1}(Z_2)) + g(F_2^{-1}(1 - Z_2))}{2}$$

Taking the expectation over both scenarios gives the desired result, $E[g(X_2)] - E[g(X_1)] < 0$.

5.4 Distribution of $D_{L_2}(j)$ and $D_{L_2}^*(j)$

The distribution for $D_{L_2}(j)$ and $D_{L_2}^*(j)$ can be described by the variances and covariances of terms in the following vector:

$$(52) \quad V_1 = [Y_{i_1}, \epsilon_M, Y_{i_1} Y_{i_2}, Y_{i_1} \epsilon_M, Y_{i_1}^2, \epsilon_M^2, Y_{i_1} Y_{i_3}, Y_{i_2} \epsilon_M]'$$

by simple calculation, the $\text{var}(V_1)$ can be described by:

$$\begin{pmatrix} \frac{1}{2}p - \frac{1}{2}p^2 & 0 & \frac{1}{2}p^2 - \frac{1}{2}p^3 - \frac{3}{4}p^4 & 0 & \frac{1}{4}p + \frac{1}{4}p^2 - \frac{1}{2}p^3 & 0 & \frac{1}{2}p^2 - \frac{1}{2}p^3 & 0 \\ 0 & \sigma^2 & 0 & p\sigma^2 & 0 & 0 & 0 & p\sigma^2 \\ \frac{1}{2}p^2 - \frac{1}{2}p^3 & 0 & \frac{1}{4}p^2 + \frac{1}{2}p^3 - \frac{3}{4}p^4 & 0 & \frac{1}{4}p^2 + \frac{1}{4}p^3 - \frac{1}{2}p^4 & 0 & \frac{1}{2}p^3 - \frac{1}{2}p^4 & 0 \\ 0 & p\sigma^2 & 0 & (\frac{1}{2}p^2 + \frac{1}{2}p)\sigma^2 & 0 & 0 & 0 & p^2\sigma^2 \\ \frac{1}{4}p - \frac{1}{4}p^2 - \frac{1}{2}p^3 & 0 & \frac{1}{4}p^2 + \frac{1}{4}p^3 - \frac{1}{2}p^4 & 0 & \frac{1}{8}p + \frac{5}{8}p^2 - \frac{1}{2}p^3 - \frac{1}{4}p^4 & 0 & \frac{1}{4}p^2 + \frac{1}{4}p^3 - \frac{1}{2}p^4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\sigma^4 & 0 & 0 \\ \frac{1}{2}p^2 - \frac{1}{2}p^3 & 0 & \frac{1}{2}p^3 - \frac{1}{2}p^4 & 0 & \frac{1}{4}p^2 + \frac{1}{4}p^3 - \frac{1}{2}p^4 & 0 & \frac{1}{4}p^2 + \frac{1}{2}p^3 - \frac{3}{4}p^4 & 0 \\ 0 & p^2\sigma^2 & 0 & (\frac{1}{2}p^2 + \frac{1}{2}p)\sigma^2 & 0 & 0 & 0 & p\sigma^2 \end{pmatrix}$$

In this section, as we focus on only a single locus, we have dropped the subscript ‘j’ from Y_{ij} , p_j and σ_j^2 . Next, by expanding the terms, and assuming $n \equiv n_M = n_R$, $\sigma_R^2 = \sigma_M^2$ and $\beta_R = \beta_M = 0$, we get

$$(53) \quad D_{L_2}(j) = \frac{-\sum_{i:R_i=1} 2Y_0Y_i}{n} - 2Y_0\epsilon_R + \frac{\sum_{i:R_i=1} \sum_{k:R_k=1} Y_iY_k}{n^2} + \frac{\sum_{i:R_i=1} 2Y_i\epsilon_R}{n} + \epsilon_R^2 - \left(\frac{-\sum_{i:M_i=1} 2Y_0Y_i}{n} - 2Y_0\epsilon_M + \frac{\sum_{i:M_i=1} \sum_{k:M_k=1} Y_iY_k}{n^2} + \frac{\sum_{i:M_i=1} 2Y_i\epsilon_M}{n} + \epsilon_M^2 \right)$$

and

$$(54) \quad \begin{aligned} \text{var}(D_{L_2}(j)) = & \left(\frac{8}{n} + \frac{4n(n-1)}{n^4} \right) \text{var}(Y_1Y_2) + \left(8 + \frac{8}{n} \right) \text{var}(Y_1\epsilon_M) + \\ & \frac{2}{n^3} \text{var}(Y_1^2) + 2\text{var}(\epsilon_M^2) + \\ & \left(-\frac{16}{n} - \frac{8}{n^2} + \frac{16}{n^2} \right) \text{cov}(Y_1Y_2, Y_1Y_3) + \\ & \left(-\frac{8}{n^3} \right) \text{cov}(Y_1Y_2, Y_1^2) + \left(-8 - \frac{-8}{n} \right) \text{cov}(Y_1\epsilon_M, Y_2\epsilon_M) \end{aligned}$$

By substituting the appropriate values from the variance matrix for V_1 , we get the variance given in section 2.3. Using the same concept of expanding the terms and straight forward calculation, we were able to find $\text{covar}(D_{1L_2}^*, D_{2L_2}^*)$.

References

- David J. Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63(3):221 – 230, 2003.
- Jennifer Couzin. Genetic privacy: Whole genome data not anonymous, challenging assumptions. *Science*, 321(5894):1268–1374, Sept 2008.
- L.A. Foreman, C. Champod, J.A. Evett, and S. Pope Lambert. Interpreting dna evidence: A review. *International Statistical Review*, 71:473–495, 2003.
- Wing K. Fung and Yue-Qing Hu. Evaluating mixed stains with contributors of different ethnic groups under the nrc-ii recommendation 4.1. *Statistics in Medicine*, 21:3583–3593, Nov 2002.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167, Aug 2008.

Mark A Jobling and Peter Gill. Encoded evidence: Dna in forensic analysis. *Nat Rev Genet*, 5:739–751, Oct 2004.

Kenneth K. Kidd, Andrew J. Pakstis, William C. Speed, Elena L. Grigorenko, Sylvester L.B. Kajuna, Nganyirwa J. Karoma, Selemani Kungulilo, Jong-Jin Kim, Ru-Band Lu, Adekunle Odunsi, Friday Okonofua, Josef Parnas, Leslie O. Schulz, Olga V. Zhukova, and Judith R. Kidd. Developing a snp panel for forensic identification of individuals. *Forensic Science International*, 164(1):20 – 32, 2006.

Stuart Macgregor, Zhen Zhen Zhao, Anjali Henders, Martin G. Nicholas, Grant W. Montgomery, and Peter M. Visscher. Highly cost-efficient genome-wide association studies using dna pools and dense snp arrays. *Nucleic Acids Res*, 36:e35, 2008.

JV Pearson, MJ Huentelman, RF Halperin, WD Tembe, S Melquist, and et al. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *American Journal of Human Genetics*, 80:126–139, 2007.

M Stoneking, D Hedgecock, R.G. Higuchi, L Vigilant, and HA Erlich. Population variation of human mtdna control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *American Journal of Human Genetics*, 48(2):370–382, 1991.

The International HapMap Consortium. The international hapmap project. *Nature* 426:789–796, December 2003.