7-2001

Outlier Detection and False Discovery Rates for Whole-genome DNA Matching

Jung-Ying Tzeng
Carnegie Mellon University

W. Byerley
University of California - Irvine

B. Devlin
University of Pittsburgh

Kathryn Roeder
Carnegie Mellon University, roeder@stat.cmu.edu

Larry Wasserman
Carnegie Mellon University, larry@stat.cmu.edu

Outlier Detection and False Discovery Rates for Whole-genome DNA Matching

Tzeng, Jung-Ying[1], Byerley, W.[2], Devlin, B.[3],
Roeder, Kathryn[1] and Wasserman, Larry[1]

[1] Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890

[2] Department of Psychiatry
University of California
Irvine CA

[3] Department of Psychiatry
Western Psychiatric Institute & Clinic
Pittsburgh, PA 15213-2593

Abstract: We define a statistic, called the *matching statistic*, for locating regions of the genome that exhibit excess similarity among cases when compared to controls. Such regions are reasonable candidates for harboring disease genes. We find the asymptotic distribution of the statistic while accounting for correlations among sampled individuals. We then use the Benjamini and Hochberg false discovery rate (FDR) method for multiple hypothesis testing to find regions of excess sharing. The p-values for each region involve estimated nuisance parameters. Under appropriate conditions, we show that the FDR method based on p-values and with estimated nuisance parameters asymptotically preserves the FDR property. Finally, we apply the method to a pilot study on schizophrenia.

Key Words: Association study, Case-control, False discovery rate with nuisance parameters, Linkage disequilibrium

1 Introduction

During the past decade scientists have had phenomenal success at discovering the genes responsible for simple genetic disorders. The success is partially due to the fact that these disorders are generally caused by one or at most a small number of defective genes and, as in Mendel's peas, these genes act in a manner that is straightforward to model. By contrast, complex disorders are those for which there is clearly a genetic basis, but the inheritance pattern is not apparent. For complex disorders, certain *alleles* (particular versions of a gene) enhance the risk of contracting the disorder but are neither necessary nor sufficient to cause the disorder. An allele associated with increased risk of disease is called a *liability allele*. Furthermore there may be many genes at various locations (*loci*) that possess liability alleles.

To discover the *liability loci* that affect the risk of human diseases, scientists exploit the fact that the DNA in a region bracketing a liability allele will tend to be passed down along with the liability allele itself from generation to generation. Through the process of *recombination* each pair of chromosomes usually breaks into a few pieces and the genetic material is exchanged between the pair. This process causes the length of the chromosomal segment shared among affected individuals to diminish, which helps to localize the position of the liability locus within a particular chromosome. Genetic linkage analysis looks for an unusually large amount of sharing of a particular chromosomal segment among the affected members of a family.

Although linkage analysis has been a powerful tool for the discovery of simple genetic disorders, it has not experienced a similar level of success for complex disorders, presumably because the power is insufficient (Risch and Merikangas 1996). To gain more power one can exploit the fact that, within a region bracketing a liability allele, the genetic material can be conserved for hundreds of generations. Analysis that looks for unusual sharing of chromosomal segments among affected members of a population, rather than among affected individuals within an extended family (pedigree), is a form of *association analysis* (e.g., McPeek and Strahs, 1999).

Typically geneticists do not initially sequence chromosomal segments. Rather they measure alleles at particular loci known as *genetic markers* at regular intervals either over the entire genome

or targeted regions of particular interest. For the sampled set of genetic markers all lying within a single chromosomal segment, the ordered string of alleles defines a *haplotype*. An excess of a certain form of haplotype in affected individuals versus unaffected individuals is consistent with the presence of a liability allele in the region defined by the haplotype. The correlation between certain alleles in a region bracketing a liability allele is called *linkage disequilibrium*. In general, the stronger the linkage disequilibrium, the easier it is to discover the approximate location of the liability allele.

For many reasons, not all affected individuals share the same haplotype in a region bracketing a liability locus: (i) many individuals will have the disorder for other reasons, either genetic or environmental; and (ii) even among those individuals whose liability traces in part to a common locus, not all will have inherited this liability allele from a common ancestor; and (iii) even for those who inherited the liability allele from a common ancestor, recombinations may have occurred within the haplotype since the introduction of the liability allele into the population. For these reasons only modest differences in haplotype frequencies are expected between the affected individuals (cases) and the unaffected individuals (controls) even if there were a liability allele in the region under investigation.

Compounding the statistical challenge, it is often unreasonable to assume that the sampled haplotypes are independent. Individuals with common ancestry are more likely to share haplotypes throughout the genome than would be predicted due to chance. In genetic studies it is expected that some individuals who share a genetic disorder also share a common ancestor, but this common ancestry is far enough in the past that it is often unknown. Furthermore, population substructure also induces correlation among individuals from the same ethnic group. Overall the data can possess a complex correlation structure that is difficult to model directly.

In this article we propose a *matching statistic* to measure the difference between the haplotype distribution of cases and controls. In our derivation of the distribution of the matching statistic we incorporate the correlation among haplotypes in a simple way. In particular we show that the distribution of the matching statistic is well approximated by its distribution assuming the haplotypes are independent and identically distributed, multiplied by a constant that reflects the perturbation

due to correlation. Because the correlation structure is not directly estimable based on a sample of haplotypes obtained from a single region of the genome, a key step in the development of the matching statistic involves demonstrating that the correlation induces an effect that is constant across the genome. Being constant, this factor is estimable, provided multiple regions of the genome have been sampled.

In genetic studies haplotypes are typically obtained for many regions (K), across the genome. Within each region a test for association is performed. There are many methods for deciding which hypotheses to reject while maintaining control over the probability of false positives at level α. For example, the well known Bonferroni method rejects a hypothesis if the p-value $P < \alpha/K$. This guarantees that the familywise error rate (FWE) – the probability of at least one false rejection – will be no larger than α. When K is large, relative to the sample size, the Bonferroni method (and other FWE controlling methods) have power tending toward 0. Benjamini and Hochberg (1995) argued that when testing many hypotheses, protecting against a single false rejection is too stringent. Instead, they suggest controlling the *false discovery rate (FDR)*, which is the fraction of false positives. For whole-genome matching, in which the number of hypotheses can be potentially very large, we find their argument compelling and so this is the approach we take.

This research was motivated by an ongoing study of schizophrenia on Palau, a remote island in Micronesia (Devlin, Roeder, Otto, Tiobech, and Byerley 2001). Schizophrenia is a complex disease that appears to have a substantial genetic basis. A noteworthy feature of the Palauan population is that it exhibits an elevated rate of schizophrenia relative to the worldwide rate. Palau has a unique history that makes it potentially amenable to gene discovery via an association study. Linguistic analyses and ethnographic studies suggest that the Palauan population developed in relative isolation, even from other Micronesian populations; nonetheless, this population shows evidence of immigration from surrounding populations (Devlin et al. 2001 and references therein). Being settled about 2000 years ago, presumably by Asian islanders, the population is both young and small in number, currently numbering 21000. Epidemics of contagious disease originating from American and European contact reduced the population to a low of 4000 about 100 years ago. These reductions enhanced the linkage disequilibrium and presumably increased the population's suitability for

an association study. In this article we illustrate the matching statistic for detecting association in a genome scan by analyzing a pilot sample of patients and controls obtained from Palau.

This paper has the following organization: Section 2 motivates the matching statistic and derives its distribution; Section 3 proves the validity of the FDR procedure for detecting outliers using the matching statistic; Section 4 describes a simulation study; Section 5 shows the results of the Palauan data analysis; and finally Section 6 presents discussion and conclusions.

2 The Matching Statistic: Quantifying Haplotype Sharing

In this section, we develop a test statistic for association, initially by ignoring correlations due to relatedness among sampled individuals (Section 2.1) and then taking the correlation into account (Section 2.2). We assume the test statistic will be computed at each of K regions of interest across the genome.

2.1 Independent Samples

Consider K regions of interest with n case and m control haplotypes sampled from each region. Each individual contributes two haplotypes to the sample. Let $H_{i(k)}$ denote the i'th sampled haplotype in region k. Assume there are R_k distinct haplotypes in segment k. Let $\pi_{al(k)} = Pr(H_{i(k)} = l)$ for a haplotype sampled from an affected individual, $i = 1, \ldots, n$ and $l = 1, \ldots, R_k$, and $\pi_{ul(k)} = Pr(H_{i(k)} = l)$ for a haplotype sampled from an unaffected individual, $i = 1, \ldots, m$ and $l = 1, \ldots, R_k$. The original data consist of two matrices, one for the cases of dimension $n \times K$, and one for the controls of dimension $m \times K$. The (i, k) entry of the case matrix is the form of the i'th haplotype at region k. The (i, k) entry of the control matrix is arranged similarly. Within each column of each matrix, assume that the haplotypes are a sample from a multinomial distribution.

An omnibus chi-square test with $R_k - 1$ degrees of freedom between column k of the cases and column k of the controls offers one possible test to determine if the haplotype distribution differs

across cases and controls at locus k. In this report we wish to investigate a statistic that tests for association using only one degree of freedom. A one-degree-of-freedom test is of interest for three reasons: (i) it has the potential of exhibiting greater power, at least in some portions of the parameter space; (ii) it is likely to achieve its asymptotic distribution with a smaller sample size; and (iii) it permits a natural extension that incorporates correlation among haplotypes. We develop these points throughout the remainder of the manuscript, and summarize them in Section 6.

To measure the degree of matching in region k, suppose we draw two case haplotypes at random. The chance they will have the same version of the haplotype is $\sum_l \pi^2_{al(k)}$. One minus this quantity is called the heterozygosity index. The heterozygosity is maximized when $\pi_{al(k)} = 1/R_k,\ l = 1, \ldots, R_k$ and minimized when $\pi_{al(k)} = 1$ for some l. If a mutation leading to increased risk of disease occurred in the population, it is most likely to be embedded within a relatively common haplotype. If a cluster of the cases traces back to this common ancestor, then the cases will have diminished heterozygosity relative to the controls. The degree of matching at locus k in the cases versus the controls can be measured by the difference in the heterozygosity indexes:

$$\mu_k = \sum_{l=1}^{R_k} \pi^2_{al(k)} - \sum_{l=1}^{R_k} \pi^2_{ul(k)}.$$

This measure tends to be large if substantial clusters of case haplotypes derive from one (or at most several) common ancestor(s), such as would be anticipated under the alternative hypothesis of association. In this article we develop a test statistic based upon this measure of association, but note that this is just one of many possible measures of association. Methods similar to those presented here could be developed for other measures as well.

Let $\hat{\pi}_{al(k)}$ be the maximum likelihood estimator for $\pi_{al(k)}$ in the cases and let $\hat{\pi}_{ul(k)}$ be the corresponding quantity for the controls. Thus $\hat{\pi}_{al(k)}$ is the observed proportion of haplotype l in the case sample at locus k and $\hat{\pi}_{ul(k)}$ is the observed proportion of haplotype l at locus in control samples. Then T_k is the maximum likelihood estimator of μ_k:

$$T_k = \sum_{l=1}^{R_k} \hat{\pi}^2_{al(k)} - \sum_{l=1}^{R_k} \hat{\pi}^2_{ul(k)}.$$

We call T_k the unstandardized matching statistic for region k.

Define $\Pi_{a(k)} = (\pi_{a1(k)}, \ldots, \pi_{aR_k(k)})$ and $\Pi_{u(k)} = (\pi_{u1(k)}, \ldots, \pi_{uR_k(k)})$ and let $\hat{\Pi}_{a(k)}$ and $\hat{\Pi}_{u(k)}$ denote the corresponding estimated quantities. The variance of T_k can be computed directly by noting that

$$\text{Var}(T_k) = \text{Var}\left(\sum_{l=1}^{R_k} \hat{\pi}^2_{al(k)} - \sum_{l=1}^{R_k} \hat{\pi}^2_{ul(k)}\right) = \text{Var}\left(\sum_{l=1}^{R_k} \hat{\pi}^2_{al(k)}\right) + \text{Var}\left(\sum_{l=1}^{R_k} \hat{\pi}^2_{ul(k)}\right)$$

and expressing $\sum_{l=1}^{R_k} \hat{\pi}^2_{al(k)}$ as a quadratic form, $\hat{\Pi}^T_{a(k)} A_{(k)} \hat{\Pi}_{a(k)}$, with $A_{(k)}$ being the R_k-dimensional identity matrix. The variance of T_k is computed in Appendix A. We call this variance, σ^2_k, the *multinomial variance* as it is computed assuming the sample of haplotypes follows the multinomial distribution. Provided $\Pi_{a(k)}$ and $\Pi_{u(k)}$ are not equal to $\frac{1}{R_k}(1, \ldots, 1)$, T_k is approximately distributed as $N(\mu_k, \sigma^2_k)$ and σ^2_k is estimable.

Most regions will not harbor a liability allele. The level of matching is assumed to be a constant value μ across these "null" regions. Because the cases may differ somewhat in their ethnic origin from the controls, μ is not assumed to be zero. However, based on genetic theory, we anticipate $\Pi_{a(k)} \approx \Pi_{u(k)}$ in null regions for any complex disease (Devlin, Roeder and Wasserman 2001) and hence μ is close to zero.

According to the association hypothesis, those regions harboring liability alleles are likely to exhibit inflated matching. The goal is to find the regions where $\mu_k > \mu$. We have now reduced the problem to the following. We have $T_k \sim N(\mu_k, \sigma^2_k)$ for $k = 1, \ldots, K$. There is a real number μ and a subset $S \subset \{1, \ldots, K\}$ such that $\mu_k = \mu$ for $k \in S$ and $\mu_k > \mu$ for $k \notin S$. The goal is to identify S^c.

2.2 Correlated Samples

In the previous section we formulated a matching statistic that measures the degree of sharing observed in each measured region of the genome. However, in so doing, we computed the multinomial variance ignoring the correlation between haplotypes due to relatedness among the individuals in our sample. In reality, the variance of the matching statistic depends upon this correlation. To address the correlation, we extend an approach pioneered by Devlin and Roeder (1999) to this setting. These authors demonstrated that when the data consists of 2×2 case-control tables computed for

each of $k = 1, \ldots, K$ biallelic markers, the variance of $\hat{\pi}_{a1(k)} - \hat{\pi}_{u1(k)}$ equals the binomial variance times a constant multiplier τ^2 that accounts for the correlation among subjects in the study. This is a useful result because one can estimate τ^2, provided K is large.

Here we demonstrate that there exists a variance in-/de-flating factor, τ, such that Var(T_k) is proportional to the multinomial variance, σ_k^2 under certain assumptions: i.e., Var$(T_k) = \tau^2 \sigma_k^2$. Thus, following Devlin and Roeder, we also can obtain the true variance by estimating τ^2 and then scaling the multinomial variance σ_k^2 by τ^2. We first provide motivation for our assumptions and then establish the result.

Recall that case (control) haplotypes are indexed $i = 1, \ldots, n$ $(i = 1, \ldots, m)$ for the sample of $n/2$ $(m/2)$ individuals, ignoring the pairing of haplotypes within an individual. Haplotype i, obtained from an affected individual, can be encoded in a binary vector of length R_k, $Y_{i(k)} = (Y_{i(k)}^1, Y_{i(k)}^2, ..., Y_{i(k)}^{R_k})$, consisting of $R_k - 1$ zeros and a single one; for example, type 2 is coded as $(0, 1, 0, ..., 0)$, and type R_k is coded as $(0, 0, ..., 0, 1)$. Let $X_{i(k)}$ denote the corresponding quantity for the i'th control haplotype. To compute the variance under the null hypothesis of no extra matching at locus k, we assume that $\Pi_{a(k)} \doteq \Pi_{u(k)} \equiv \Pi_{(k)}$, for any $k \in S$. Consequently, under the null hypothesis, $Y_{i(k)} = \left(Y_{i(k)}^1, Y_{i(k)}^2, ..., Y_{i(k)}^{R_k}\right)$ and $X_{i(k)} = \left(X_{i(k)}^1, X_{i(k)}^2, ..., X_{i(k)}^{R_k}\right)$ identically, but not independently, follow a Multinomial distribution with sample size one and probability vector $\Pi_{(k)}$. For a fixed observation i, let $\rho_{(k)}^{lh}$ denote the usual multinomial correlation between two forms of a haplotype (type l and type h), i.e.,

$$\text{Corr}(Y_{i(k)}^l, Y_{i(k)}^h) \;=\; \text{Corr}(X_{i(k)}^l, X_{i(k)}^h) \;=\; \rho_{(k)}^{lh} = -\sqrt{\frac{\pi_{l(k)}\pi_{h(k)}}{(1-\pi_{l(k)})(1-\pi_{h(k)})}}.$$

Notice that $\rho_{(k)}^{lh}$ is independent of i but depends on the region k, and is the same for cases and controls under the null hypothesis.

For haplotype form l, let $f_{ij}^{(Y)}$ denote the correlation between two case haplotypes, $f_{ij}^{(X)}$ the correlation between two control haplotypes, and $f_{ij}^{(XY)}$ the correlation between a case and a control

haplotype. We assume these do not depend on k and l, i.e.

$$\begin{aligned} \text{Corr}(Y^l_{i(k)}, Y^l_{j(k)}) &= f^{(Y)}_{ij} \\ \text{Corr}(X^l_{i(k)}, X^l_{j(k)}) &= f^{(X)}_{ij} \quad . \\ \text{Corr}(Y^l_{i(k)}, X^l_{j(k)}) &= f^{(XY)}_{ij} \end{aligned} \tag{1}$$

Next consider the pairwise correlation between different forms of haplotypes ($l \neq h$) and different measured haplotypes ($i \neq j$). Assume the correlation coefficients can be expressed as:

$$\begin{aligned} \text{Corr}(Y^l_{i(k)}, Y^h_{j(k)}) &\equiv {f^{lh}_{ij(k)}}^{(Y)} = \rho^{lh}_{(k)} f^{(Y)}_{ij} \\ \text{Corr}(X^l_{i(k)}, X^h_{j(k)}) &\equiv {f^{lh}_{ij(k)}}^{(X)} = \rho^{lh}_{(k)} f^{(X)}_{ij} \\ \text{Corr}(Y^l_{i(k)}, X^h_{j(k)}) &\equiv {f^{lh}_{ij(k)}}^{(XY)} = \rho^{lh}_{(k)} f^{(XY)}_{ij}. \end{aligned} \tag{2}$$

Genetic theory supports these assumptions. The correlation f_{ij} in (1) is assumed to be independent of k and l because it is determined by the ancestry of the chromosomal segments themselves and is not a function of the haplotype form l or the haplotype segment k. Indeed, based upon genetic theory the correlation between two individuals f_{ij} is equal to the probability two chromosomal segments are inherited from a common ancestor, perhaps many generations in the past. Extant chromosomal segments deriving from a common ancestor are said to be *identical by descent* (ibd). The term "ibd" emphasizes that the two chromosomal segments match because they are from a common ancestor rather than matching due to chance. Based on this we obtain

$$\begin{aligned} \text{Cov}(Y^l_{i(k)}, Y^l_{j(k)}) &= \text{E}(Y^l_{i(k)} Y^l_{j(k)}) - \text{E}(Y^l_{i(k)})\text{E}(Y^l_{j(k)}) \\ &= \Pr(Y^l_{j(k)} = 1 \mid Y^l_{i(k)} = 1)\, \Pr(Y^l_{i(k)} = 1) - \Pr(Y^l_{i(k)} = 1)\Pr(Y^l_{j(k)} = 1) \\ &= \left[f_{ij} + (1 - f_{ij})\pi_{l(k)} \right] \pi_{l(k)} - \pi^2_{l(k)} \\ &= f_{ij}\, \pi_{l(k)}(1 - \pi_{l(k)}). \end{aligned}$$

From this it follows that $\text{Corr}(Y^l_{i(k)}, Y^l_{j(k)}) = f_{ij}$. By the same manner, we obtain $\text{Cov}(Y^l_{i(k)}, Y^h_{j(k)}) = \left(-\pi_{l(k)}\pi_{h(k)}\right) f_{ij}$. That is, $\text{Corr}(Y^l_{i(k)}, Y^h_{j(k)}) = \rho^{lh}_{(k)} \cdot f_{ij}$.

THEOREM 1: *Let*

$$\tau^2 = 1 + \tfrac{2mn}{m+n} \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j>i} f^{(Y)}_{ij} + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j>i} f^{(X)}_{ij} - \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} f^{(XY)}_{ij} \right), \tag{3}$$

and define σ_k^2 to be the variance of T_k obtained assuming the $n+m$ haplotypes are independent and identically distributed from a multinomial $(\Pi_{(k)})$ distribution. If the correlation structure implied by (1) and (2) holds and $m/n \to \psi$ for some positive constant ψ, then

$$\lim_{n,m\to\infty} \textit{Var}\left[\frac{T_k}{\tau\sigma_k}\right] = 1. \tag{4}$$

PROOF: Because the analytical form of the exact variance of correlated T_k is intractable, we study the relationship between the variance of the correlated sample and the variance of an independent and identically distributed multinomial sample via the delta method approximation. By the delta method, the multinomial variance is approximately equal to:

$$\sigma_k^2 \doteq \left(\frac{1}{n}+\frac{1}{m}\right)\left[4\sum_{l=1}^{R_k}\pi_{l(k)}^3(1-\pi_{l(k)}) - 8\sum_{l=1}^{R_k}\sum_{h>l}\pi_{l(k)}^2\pi_{h(k)}^2\right].$$

Similarly, the approximate Var(T_k), based on a correlated sample, is approximately equal to

$$\begin{aligned}
&\sigma_k^2 + \text{Cov}\left[\sum_{l=1}^{R_k}\hat{\pi}_{al(k)}^2,\ \sum_{l=1}^{R_k}\hat{\pi}_{ul(k)}^2\right] \\
= \ &\sigma_k^2 + 4\sum_{l=1}^{R_k}\pi_{l(k)}^3(1-\pi_{l(k)})\cdot 2\left({}^{1}\tfrac{}{n^2}\sum_{i=1}^{n}\sum_{j>i}f_{ij}^{(Y)} + \tfrac{1}{m^2}\sum_{i=1}^{m}\sum_{j>i}f_{ij}^{(X)} - \tfrac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}f_{ij}^{(XY)}\right) \\
&- 8\sum_{l=1}^{R_k}\sum_{h>l}\pi_{l(k)}^2\pi_{h(k)}^2\cdot\tfrac{1}{\rho_{(k)}^{lh}}\cdot 2\left(\tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{j>i}f_{ij(k)}^{lh\ (Y)} + \tfrac{1}{m^2}\sum_{i=1}^{m}\sum_{j>i}f_{ij(k)}^{lh\ (X)} - \tfrac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}f_{ij(k)}^{lh\ (XY)}\right).
\end{aligned}$$

Plugging in the quantities given in (1) and (2), the result follows. $\diamond$

So far we verified the variance factorization (4) and can approximate the distribution of the matching statistic $Z_k = (T_k-\mu)/(\tau\sigma_k)$ under the null hypothesis with a standard normal. With the analytic formula of σ_k^2, the only remaining requirements in standardization are estimates of μ and τ. Notice τ^2 does not depend upon k, that is, no matter which region we are considering, the correlation among sampled haplotypes affects the variance of T_k multiplicatively in the same way. The most important consequence of this result is that τ can be estimated from the sample of T_k's.

A U-statistic can be used to estimate τ. For $1 \le k < g \le K$, define

$$U_{kg} = \frac{T_k - T_g}{\sqrt{\sigma_k^2+\sigma_g^2}}.$$

Because $T_k \approx N(\mu, \tau^2\sigma_k^2)$, it follows that for pairs of null loci, $U_{kg} \approx N(0, \tau^2)$. This suggests using the sample variance of the U_{kg} as an estimate of τ^2 which is thus a U-statistic since the U_{kg} depend on the pairs (T_k, T_g). Alternatively, one can use a robust scale estimator applied to $\{U_{kg}\}$ such as $\hat{\tau} = \text{median}(\mid U_{kg} \mid)/0.65$. Similarly, μ can be estimated with either the mean or median of the T_k's. When estimating μ and τ, it is preferable to use robust estimators because μ and τ reflect quantities defined for the null loci.

3 Outliers and False Discovery Rates

3.1 False Discovery Rate

We will identify the outliers by testing $H_{0k} : \mu_k = \mu$ versus $H_{1k} : \mu_k > \mu$ for $k = 1, \ldots, K$. To correct for multiple testing we will use the FDR (false discovery rate) method of Benjamini and Hochberg (1995). We begin this section by reviewing their method. Then we show how the method can be used in our setting.

Consider testing a set of null hypotheses $H_{01}, \ldots, H_{0K}$. Let $P_1, \ldots, P_K$ be the p-values associated with the K tests. Suppose we reject some subset of these hypotheses. The realized false discovery rate Q is defined to be the number of false rejections divided by the total number of rejections. Q is taken to be 0 if no hypotheses are rejected. The Benjamini and Hochberg method for controlling $E(Q)$ is as follows. First, order the p-values, $P_{(1)} \leq \cdots \leq P_{(K)}$ and then define $d = \max\{j : P_{(j)} < j\alpha/K\}$. Finally, reject all hypotheses whose p-values are less than or equal to $P_{(d)}$. Benjamini and Hochberg proved that this procedure ensures that $E(Q) \leq \alpha$, no matter how many nulls are false and no matter what is the distribution of the p-values when the null is false. The operating characteristics of the method are discussed in Genovese and Wasserman (2001). In particular, the method has much higher power than Bonferroni when K is large.

This method can be adapted to our setting. Let $T_1, \ldots, T_K$ be test statistics where $T_k \approx N(\mu_k, \sigma_k^2\tau^2)$. For the moment, suppose that μ, τ, σ_k^2 are known and that T_k has exactly a Normal distribution. The Benjamini and Hochberg procedure can be applied to test $H_{0k} : \mu_k = \mu$ using the usual p-values

defined by $P_k = 1 - \Phi(Z_k)$ where $Z_k = (T_k - \mu)/(\sigma_k \tau)$ and Φ is the standard normal cumulative distribution function.

In our setting there are several complications. First, T_k is only asymptotically Normal in the number of cases and controls. For large n we have found (by simulation) that this is not a serious problem; see also the example in Section 5. Second, the T_k may be correlated. Benjamini and Yekutieli (2001) showed that the inequality $E(Q) \leq \alpha$ still holds for correlated tests if d is replaced by $\tilde{d} = \max\{j : P_{(j)} < j\alpha/c_K K\}$ where $c_K = \sum_{k=1}^{K}(1/K) \approx \log K$. This leads to a more conservative procedure. However, our experience is that the types of correlation in our problem are mild and do not inflate the FDR. Thus, we have ignored the correction. Finally, μ, τ and σ_k^2 are unknown and must be estimated. We will now show that, at least asymptotically, the FDR property is preserved even when the p-value is estimated by inserting consistent estimates of these parameters.

3.2 FDR With Nuisance Parameters

For simplicity, we take $\sigma_k = 1$ and known in what follows and concentrate on μ and τ. The extension to unknown σ_k is straightforward, albeit tedious. Let $S(z) = 1 - \Phi(z)$. The p-value defined above for testing $H_{0k} : \mu_k = \mu$ can be written as $P_k = S(Z_k)$, where $Z_k = (T_k - \mu)/\tau$. Define the estimated p-value by $\hat{P}_k = S(\hat{Z}_k)$ where $\hat{\mu}$ and $\hat{\tau}$ are the estimates of μ and τ, respectively, and $\hat{Z}_k = (T_k - \hat{\mu})/\hat{\tau}$.

Consider the FDR procedure based on the estimated p-values in the place of the true p-values. In what follows, we assume that when computing $\hat{P}_k$, the estimates $\hat{\mu}$ and $\hat{\tau}$ are Studentized, i.e. the k^{th} observation is omitted. We denote these estimates by $\hat{\mu}_{(k)}$ and $\hat{\tau}_{(k)}$. Let F_k denote the common cumulative distribution function of $\hat{P}_k$ under the null.

THEOREM 2. *For every* $K \geq 1$,

$$E(Q) \leq \alpha \sup_{\frac{\alpha}{K} \leq p \leq \alpha} \frac{F_k(p)}{p}.$$

THEOREM 3. For any fixed α,

$$\limsup_{K\to\infty} E_K(Q) \leq \alpha.$$

REMARK: The estimators in Section 3 are asymptotically biased, as is the case in general for both robust and non-robust estimators in the presence of outliers. However, if we let the fraction of outliers tend to 0 as K grows, this bias disappears asymptotically. This assumption is realistic because it reflects the fact that the fraction of liability genes is small relative to the size of the genome.

Proofs for these results can be found in Appendix B.

4 Simulation

A simulation study was conducted to investigate the FDR and the power of the procedure under various settings. In each experiment we used the robust estimators of μ and τ. We generated data from $K = 100$ regions with sample size $n = m = 500$ (or 250 individual cases and controls, each with a pair of haplotypes). The number of distinct haplotypes was set at $R_k = 32$ and the nominal level of significance was set at 0.05. To obtain the average performance of the procedure, 1000 datasets were generated for each configuration under investigation. Power, in this case, is defined as the average fraction of alternative hypotheses rejected using the FDR method.

To validate the theorems concerning the detection of outliers, we generated data under the null hypothesis in three different ways. To investigate the procedure under the simplest setting we set $\Pi_u = \Pi_a = \Pi^0$, in which the fixed haplotype distribution Π^0 was chosen to have four levels of haplotype probabilities $(0.0125, 0.025, 0.0375, 0.05)$, with eight consecutive repetitions of each level to obtain a total of $R_k = 32$ types. We set $\alpha = 0.05$ in the FDR procedure. For this setting the mean FDR was 0.026. Next we allowed $\Pi_{u(k)}$ to vary randomly as a function of k by sampling $\Pi_{u(k)}$ from the $\mathrm{Dirichlet}(32 \times \Pi^0)$. By setting the haplotype frequencies of cases equal to the controls,

$\Pi_{a(k)} \equiv \Pi_{u(k)}$, we fixed $\mu = 0$. For this setting the mean FDR was 0.046. Finally, the model allows for general location shifts, but Π_u is assumed to be equal to Π_a when computing the variance. We investigated the robustness of the procedure to this approximation when $\mu \neq 0$, This time we sampled both $\Pi_{u(k)}$ and $\Pi_{a(k)}$ from the $\mathrm{Dirichlet}(32 \times \Pi^0)$. The choice of 32 corresponds to at least as much variation as is likely to be observed between case and control populations, under the null hypothesis (Devlin et al. 2001). For this setting the mean FDR was 0.010. We conclude that the procedure performs well under the null hypothesis. Indeed, it is somewhat conservative.

To investigate the power under various types of outlier models we preset $\Pi_{u(k)}$ at Π^0 for all the null loci and then perturb this distribution for a set of alternative loci. We considered two different levels of contamination: $H = 5$ outliers and $H = 20$ outliers. For the remaining regions, $k = H + 1, \ldots, 100$, $\Pi_{a(k)} \equiv \Pi^0$ as defined above for the controls. We perturb Π^0 in four ways to obtain $\Pi_{a(k)q}, k = 1, \ldots, H$, $q = 1, \ldots, 4$: $\Pi_{a(k)q} = (1 - a)\Pi^0 + a\ \mathrm{spike}_q$. Let spike_q $q = 1, \ldots, 4$ denote probability vectors of length R_k. Let spike_1 be a point mass at $l = 1$, spike_2 be a point mass at $l = 25$, spike_3 be a mass of 0.5 at $l = 1$ and 9, spike_4 be a mass of 0.5 at $l = 17$ and 25 with $a = 0.15$ (Figure 1). The first and second conditions simulate the performance when a fraction a of the haplotypes trace back to a single ancestral haplotype and the third and fourth conditions simulate the performance when a fraction $a/2$ of the haplotypes trace back to one of a pair of ancestral haplotypes.

The size of the deviation from the null, as measured by $\mu_{kq}, k = 1, \ldots, H,\ q = 1, \ldots, 4$, depends greatly upon the relative frequency of the haplotype that is associated with the disease. For $q = 1, \ldots, 4$, μ_{kq} equals 0.015, 0.025, 0.006 and 0.012, respectively. Clearly the biggest deviation occurs when the associated haplotype is also common in the controls and the least detectable deviation occurs when two haplotypes are associated with the disease and these haplotypes are relatively rare in the controls. Not surprisingly the latter condition ($q = 3$) exhibits considerably lower power than the other three conditions (Table 1). Overall, the relative power is correlated with the size of the true deviation, μ_k, but the power also depends upon σ_k. For instance, $\mu_1 \approx \mu_4$, but the power is considerably less for the latter configuration, because the variance is larger for this haplotype frequency distribution.

The number of regions deviating from the null, H, also affects the overall performance of the method (Table 1). When 20% of the regions deviate from the null both $\hat{\mu}$ and $\hat{\tau}$ are positively biased and the bias in $\hat{\tau}$ is substantial. Although this bias deflates the power, the FDR rate is maintained at a conservative level (considerably less than 0.05, the nominal level) for all 8 scenarios investigated. The procedure is more conservative than the nominal level due to the bias in the estimates of μ and τ. In most practical settings H is a very small fraction of K and consequently the bias will be considerably less than observed for this simulation.

To obtain a sense of how powerful the matching statistic is relative to competing methods we compared it to the omnibus chi-square test with $R_k - 1$ degrees of freedom (Table 2) using Holm's correction (Holm 1979) for both methods. The null distribution of the goodness-of-fit statistic was obtained using a permutation test. This test is only a valid competitor when $\tau = 1$. With the exception of condition $q = 3$, the matching statistic was either more powerful or roughly equivalent to the goodness-of-fit statistic in performance.

In the four simulated conditions investigated thus far, none of the alleles dominate in frequency under the null hypothesis. To complete our investigation we also considered a scenario with two common alleles ($\pi_1^* = \pi_2^* = 0.155$) and three types of rare alleles (0.016,0.023,0.03), each with ten copies, so that $R_k = 32$. Under the alternative hypothesis $\Pi_{a(k)} = (1 - a)\Pi^* + a \text{ spike}_1$. In this setting the matching test clearly dominates the goodness-of-fit statistic with power equal to 67.3% vs. 24.6%.

The general principle appears to be that the matching statistic is more powerful when the associated haplotype(s) is (are) relatively common in the control population. When the associated haplotype(s) is (are) relatively rare, then μ_k tends to be near zero and the goodness-of-fit test is more powerful. In fact, under some conditions $\mu_k \approx 0$ and the matching statistic has power equal to the size of the test. The relative strength of the matching statistic to the goodness-of-fit statistic is also greater when R_k is large.

It is also worth noting that the size of the matching test in all comparisons was smaller than the nominal size of the test. Apparently μ and τ are estimated with some bias, which leads to a conservative test.

5 Data Analysis

To illustrate the proposed methods we analyze a small sample of schizophrenia patients and controls sampled from the island nation of Palau. As part of an ongoing linkage study of schizophrenia on Palau, seven extended pedigrees have been ascertained and genotyped. We utilize a portion of these data for a pilot study of association. From the extended pedigrees, 22 cases and 27 controls were selected for further analysis. In a future study we anticipate supplementing our sample by obtaining a larger sample of both cases and controls.

Schizophrenia is a mental illness characterized by disordered thoughts, behaviors and language. Its identification is through a collection of "positive" symptoms together with "negative" symptoms. The positive symptoms include hallucinations (e.g., hearing voices or seeing things that do not exist), delusions (e.g., holding false beliefs, such as that one is being watched, spied upon, or plotted against), and disorganization or incoherence of thought or speech; the "negative" symptoms are, for example, lack of normal emotional response, withdrawal from others, neglect of grooming and hygiene, and poor work performance.

By the time the illness is diagnosed brain structure and chemistry have been altered. Ultimate causes appear to involve both acquired and genetic factors. We concentrate on the genetic factors, attempting to identify variants in genes that generate higher risk for schizophrenia. To date the scientific community has made only limited progress toward identification of the genetic basis for this challenging complex disease.

A remote island nation in Micronesia, Palau covers an archipelago of more than 200 islands scattered over 125 miles of the South Pacific. The islands lie 600 miles north of New Guinea and 550 miles east of the Phillipines. Palau exhibits a slightly elevated rate of schizophrenia, 2.77% in males and 1.24% in female, compared to the sex-averaged estimate of 0.5-1% worldwide.

Carbon dating (Takayama, 1981) suggests Palau was first populated about 2000 years ago and the initial population size was small, probably less than 50. The population grew to 20,000 by 227 years ago but decreased to 4,000 about 100 years ago because of disease epidemics. The current population size is around 21,000. Our and other results suggest that the Palauan population has

developed in isolation, compared to say European populations, but it experienced a surprising level of immigration for such a remote region (Simmons et al, 1965; Devlin et al, 2001). Genetic analysis suggests the original population appears to have migrated from island Southeast Asia, with some later migrations from Melanesia. Studies on Palau (Devlin et al., 2001) indicate substantial linkage disequilibrium (i.e., haplotype-sharing) exists in this population, making it ideal for an association study.

The population history of Palau facilitates the search for schizophrenia liability genes. First, the linkage disequilibrium for Palauan people is enlarged by the recent population bottleneck and extends potentially as far as 10 to 20 cM (Devlin et al., 2001). Second, the isolation of Palau makes it easier to detect any schizophrenia genes introduced by recent migration. This is because these foreign chromosomes will be relatively prominent compared to the general Palauan chromosomes. Based on these facts, we believe it is instructive to search for association even with the 10cM marker grid available from the Palauan linkage study. Nevertheless, this 10cM grid is much sparser than generally required for detecting association between markers and disease genes (Ott, 2000). In followup studies we hope to have a denser grid of markers.

Due to their population history, we assume most Palau natives are at least distantly related. Of the 22 cases in this study, some are close relatives, such as cousins, and other pairs are not obviously related. For controls, we chose 27 people who were either not known to be closely related to schizophrenics or were from a pedigree containing schizophrenics, but were themselves less likely to carry a disease allele. As a group, the controls are not as closely related to one another as are the cases. Nevertheless, some of the controls are closely related to some of the cases. Overall, the genetic relationships among study subjects generates a complex correlation structure among the sampled haplotypes. This is clearly not a standard case-control study, and we anticipate the correlation among haplotypes to have some effect on the variance of the test statistic.

To explore the relationship between familial relationship and matching of genetic material, all cases with a known degree of familial relationship were compared. Figure 2 shows the relationship between the degree of relatedness and the level of haplotype similarity for these case pairs. The Y-axis indicates the degree of relatedness between two cases; the lower the value, the more closely

they are related to each other. The X-axis is the average proportion of markers with matching allele types; the average is taken over 4 possible combinations since each person has two chromosomes. The plot shows that as two people are more closely related to each other, they share more common genetic material. The correlation coefficient is extremely high ($\mathrm{corr} = -0.85$).

A critical selection criterion for the 22 cases and 27 controls was the ability to determine unambiguous haplotypes for these individuals, which were obtained using the linkage program Simwalk2 (Sobel and Lange 1996). Thus for each individual, we have haplotypes for 22 pairs of autosomal chromosomes, with 37 markers on the largest chromosome (Chromosome 1), descending to 8 markers on the smallest chromosome (Chromosome 22). The genetic markers are STRs and the allele type represents the number of repeats observed at a particular locus. Because chromosomes occur in pairs, each individual contributes two haplotypes to the dataset. Table 3 displays a portion of the haplotype data for Chromosome 22. Each row records the alleles of markers on a chromosome. For example, from the last row of Table 3, the 27th control has the haplotype (5,8,7,2,3,11,3,5) on one of its 22nd chromosomes.

We defined regions using a moving bin encompassing adjacent pairs of markers. With this definition we obtained $K = 453$ regions. The haplotype frequency distribution varied by region, but $R_k \approx 32$ and one or more forms were generally considerably more common than the others.

From these 453 observations we obtain $\hat{\mu} = 0.007$ and $\hat{\tau} = 0.9046$. Surprisingly, even though we had anticipated $\tau > 1$ due to the strong positive correlation between cases, two factors offset this expectation: the cases are also correlated with the controls which reduces the variance of T_k; and for small samples $\hat{\sigma}_k$ has a slight positive bias. To compensate, $\hat{\tau}$ has a slight negative bias. We observed the same phenomenon in simulated data for small samples (results not shown). This phenomenon did not inflate FDR in our simulations.

To evaluate the assumption of normality we examine the normal scores plot of the standardized matching statistics Z_k (Figure 3). The tests statistics show a surprising degree of consistency with a normal distribution considering the size of the data set. From this figure it is also clear that none of the statistics appear to be unusually large relative to the remainder of the sample. In fact, none of the regions indicates a significant association with the disease using either the FDR or a Bonferroni

procedure.

Plotting the test statistics as a function of the approximate relative location of the haplotypes on the 22 autosomal chromosomes indicates several statistics tend to approach significance in regions that have shown promising signals in other studies of schizophrenia (Figure 4). Considering the size of the sample and the coarseness of the marker grid, the test undoubtedly has low power. Followup studies will have better power.

6 Discussion

In this article we define a statistic, called the *matching statistic*, for locating regions of the genome that exhibit excess similarity of haplotypes within case haplotypes relative to the controls. This statistic is of interest because it identifies regions that are reasonable candidates for locating disease genes. It is a practical alternative to a statistic developed in Devlin, Roeder and Wasserman (2000), which tested for excess ibd sharing assuming an extremely dense grid of genetic markers.

In many case-control association studies the sampled haplotypes are correlated either because the subjects are obviously related, cryptically related (related, but the relationship is unknown) or related due to common ethnic background. We find the asymptotic distribution of the matching statistic while accounting for correlations among sampled individuals. The approach taken here could potentially be extended to many other statistics that measure haplotype sharing or, more generally linkage disequilibrium. In fact, the performance of the matching statistic depends very strongly on the way in which the distribution of the case haplotypes deviates from the distribution of the control haplotypes. It would be interesting to investigate the performance of other statistics sensitive to linkage disequilibrium as well.

In motivating a one degree-of-freedom test we noted that such a statistic was likely to have greater power for some alternatives than an omnibus test; in Section 4 we identify the type of alternatives for which the statistic obtains a competitive advantage and it appears that this type of alternative arises frequently in practice. Naturally there are alternative haplotype distributions for which the omnibus test has greater power than the matching statistic. The advantage of the matching

statistic is that it achieves its asymptotic distribution for a modest sample size, hence it is amenable to corrections for correlated samples such as the one described in Section 2.2. Although in principle the goodness-of-fit test could also be corrected in a similar fashion, this test does not achieve its asymptotic distribution for moderate sized samples if R_k is large.

In our treatment of correlation among haplotypes we ignore the known relationships among subjects, allowing τ to adjust for correlations due to known familial relationships as well as unknown relationships. It is likely that greater power could be obtained if the known relationships were modeled overtly in a manner such as that described by Slager and Schaid (2001).

Determining which regions in the genome exhibit significant association involves a large number of hypothesis tests. In problems of this nature cogent arguments can be made that it is more appropriate to control the FDR rather than the FWE because the FDR method offers a more powerful option for discovering genomic regions that potentially possess liability alleles. In its formulation the Benjamini and Hochberg procedure controls the FDR for given a set of independent p-values. In our application the p-values for each region involve estimated nuisance parameters. We show that under appropriate conditions, the FDR method based on p-values with estimated nuisance parameters, asymptotically preserves the FDR property. These results should be useful for other applications as well.

Appendix A: Variance of T_k

We compute the variance of T_k under the assumption of iid haplotypes, and hence $n\hat{\Pi}_{a(k)}$ is distributed multinomial $(n; \Pi_{a(k)})$ and $m\hat{\Pi}_{u(k)}$ is distributed multinomial $(m; \Pi_{u(k)})$. We suppress the subscript (k) to simplify the notation.

Here we express $\sum_{l=1}^{R} \hat{\pi}_{al}^2$ in a quadratic form, $\hat{\Pi}_a^T A \hat{\Pi}_a$, with A being the R-dim identity matrix. Without extra complexity we obtain the general variance formula of the quadratic form $\hat{\Pi}_a^T A \hat{\Pi}_a - \hat{\Pi}_u^T A \hat{\Pi}_u$:

Define

$$\Sigma_n = \frac{1}{n}\left(\text{Diag}(\Pi_a) - \Pi_a \Pi_a^T\right),$$

where $\text{Diag}(\Pi_a)$ is the diagonal matrix of $(\pi_{a1}, \ldots, \pi_{aL})$, $\text{Diag}(\Pi_a^{(2)})$ is the diagonal matrix of $(\pi_{a1}^2, \ldots, \pi_{aL}^2)$ and ${\Pi_a^{(2)}}^T = (\ \pi_{a1}^2,\ \pi_{a2}^2,\ \ldots,\ \pi_{aR}^2\ )$.

Let $A = [a_{ij}]$, $(dA)^T = (\ a_{11},\ a_{22},\ ...,\ a_{RR}\ )$, $A^{(2)} = [a_{ij}^2]$, and ${(dA^{(2)})}^T = (\ a_{11}^2,\ a_{22}^2,\ ...,\ a_{RR}^2\ )$. Define $(d0A)$ as A with the diagonal elements zeroed out, A_{-l} = matrix $d0A$ with the lth-row and the lth-column deleted,

$$B \ = \ \begin{bmatrix} a_{11} & a_{22} & ... & a_{RR} \\ a_{11} & a_{22} & ... & a_{RR} \\ \vdots & \vdots & & \vdots \\ a_{11} & a_{22} & ... & a_{RR} \end{bmatrix},$$

and ${E_a}^T = (\ a_{11}\pi_{a1},\ a_{22}\pi_{a2},\ ...,\ a_{RR}\pi_{aR}\ )$.

Now Var $\left(\hat{\Pi}_a^T A \hat{\Pi}_a\right)$ equals

$$\begin{aligned}
& \tfrac{(n-1)(n-2)(n-3)}{n^3} \times [\ \Pi_a^T A \Pi_a\ ]^2 \\
+\ & \tfrac{(n-1)(n-2)}{n^3} \times \\
& \{\quad 2 \times \text{sum of all elements of } \left[\sum_{l=1}^{R} \pi_{al} a_{ll} \cdot A_{-l}\right] \\
& \ +\ 4 \times \text{sum of off-diagonal elements of } [\text{Diag}(\Pi_a)\ A\ \text{Diag}(\Pi_a)\ A\ \text{Diag}(\Pi_a)] \\
& \ +\ 4 \times \text{tr}\left[(d0A)\ \text{Diag}(\Pi_a)\ B\ \text{Diag}(\Pi_a^{(2)})\right] \\
& \ +\ 2 \times {\Pi_a}^T\ (dA)\ (dA)^T\ \Pi_a^{(2)}
\end{aligned}$$

$$
\begin{aligned}
&+\ 4 \times {\Pi_a}^T\ A^{(2)}\ \Pi_a^{(2)}\ \} \\
&+\ \tfrac{(n-1)}{n^3} \times \left\{\ 4 \times {E_a}^T\ A\ \Pi_a\ +\ {\Pi_a}^T\ (dA)\ (dA)^T\ \Pi_a\ +\ 2 \times {\Pi_a}^T A^{(2)}\ \Pi_a\ \right\} \\
&+\ \tfrac{1}{n^3} \times \left(dA^{(2)}\right)^T\ \Pi_a \\
&-\ \left[\ \mathrm{tr}[\ A\ \Sigma_n\ ]\ +\ {\Pi_a}^T\ A\ \Pi_a\ \right]^2 .
\end{aligned}
$$

Using the same general form Var $\left(\hat{\Pi}_u^T A \hat{\Pi}_u\right)$ is obtained.

Appendix B: Proof of Theorems 2 and 3.

The proof of Theorems 2 and 3 require a few lemmas. Let

$$R_k = \frac{\tau}{\hat{\tau}_{(k)}} \ \text{ and } \ D_k = \mu - \hat{\mu}_{(k)},$$

and let G_k denote the (common) joint distribution of (D_k, R_k) under the null.

LEMMA 1. *The cdf F_k is given by*

$$F_k(p) = \int\int S\left(\frac{S^{-1}(p)}{r} - d\right) dG_k(r,d).$$

PROOF. Observe that

$$\hat{Z}_k = \frac{T_k - \hat{\mu}_{(k)}}{\hat{\tau}_{(k)}} = R_k Z_k + R_k D_k.$$

Hence, $\hat{P}_k = S(\hat{Z}_k) = S(R_k Z_k + R_k D_k)$. Also, note that Z_k is independent of (D_k, R_k). Hence,

$$
\begin{aligned}
F_k(p) &= Pr(\hat{P}_k < p) \\
&= Pr(S(R_k Z_k + R_k D_k)) < p) \\
&= Pr(R_k Z_k + R_k D_k > S^{-1}(p)) \\
&= Pr\left(Z_k > \frac{S^{-1}(p)}{R_k} - D_k\right) \\
&= \int\int Pr\left(Z_k > \frac{S^{-1}(p)}{r} - d\right) dG_k(r,d) \\
&= \int\int S\left(\frac{S^{-1}(p)}{r} - d\right) dG_k(r,d).\ \diamond
\end{aligned}
$$

In what follows, we will make use of the following well known relations based on Mills' ratios: $S(z) \leq \phi(z)/z$ and, for all $0 < p < \alpha$, $q(p) = [2\log(1/p) - \log\log(1/p) - r(p)]^{1/2}$ where $0 \leq r(p) \leq \gamma$ and γ is a constant that depends only on α. It follows from these relations that, for $c > 0$,

$$\phi(q(p)c) \leq Cp^{c^2}\left(\log(1/p)\right)^{c^2/2} \tag{5}$$

where $C = (2\pi)^{-1/2}e^{c^2\gamma/2}$. In general, C will denote a generic positive constant, possibly with different values in different expressions but not depending on p or K.

LEMMA 2. *For any $\alpha > 0$,*

$$\sup_{\frac{\alpha}{K} \leq p \leq \alpha} \left| \frac{F_k(p)}{p} - 1 \right| \to 0 \tag{6}$$

as $K \to \infty$.

PROOF. Let $W_k = R_k^{-1}$ and define $h(d, w, p) = S(q(p)w - d)$ where $q(p) = S^{-1}(p)$. Hence, $F_k(p) = Eh(D_k, W_k, p)$. Expanding $h(d, w, p)$ around $(d, w) = (0, 1)$ yields

$$\frac{h(D_k, W_k, p)}{p} = 1 + \frac{D_k\phi(q(p)W_k - \tilde{D}_k)}{p} - \frac{(W_k - 1)q(p)\phi(q(p)\tilde{W}_k - D_k)}{p}$$

where ϕ is the standard normal density, $\tilde{D}$ and $\tilde{W}$ depend on p and satisfy $0 \leq |\tilde{D}| \leq |D|$ and $|1 - \tilde{W}| \leq |1 - W|$. Thus,

$$\left| \frac{h(D_k, W_k, p)}{p} - 1 \right| \leq \frac{|D_k|\phi(q(p)W_k - \tilde{D}_k)}{p} + \frac{|W_k - 1|q(p)\phi(q(p)\tilde{W}_k - D_k)}{p}. \tag{7}$$

Using (5), for $\alpha/K \leq p \leq \alpha$, we have that

$$\begin{aligned}
\frac{|D_k|\phi(q(p)W_k - \tilde{D}_k)}{p} &\leq |D_k|\frac{\phi(q(p)W_k)}{p}\exp\left\{2q(p)W_k|\tilde{D}_k|\right\} \\
&\leq |D_k|\frac{\phi\left(q\left(\frac{\alpha}{K}\right)W_k\right)}{\frac{\alpha}{K}}\exp\left\{2q\left(\frac{\alpha}{K}\right)W_k|D_k|\right\} \\
&\leq C\alpha^{W_k^2-1}|D_k|K^{1-W_k^2}\left(\log(K/\alpha)\right)^{W_k^2/2} \\
&\quad \times \exp\left\{2\sqrt{2\log(K/\alpha)}W_k|D_k|\right\} \equiv U_K.
\end{aligned} \tag{8}$$

By a similar argument,

$$\begin{aligned}
\frac{|W_k - 1|q(p)\phi(q(p)\tilde{W}_k - D_k)}{p} &\leq C\alpha^{W_k^2-1}|W_k - 1|\sqrt{\log(K/\alpha)}K^{1-W_k^2}\left(\log(K/\alpha)\right)^{W_k^2/2} \\
&\quad \times \exp\left\{2\sqrt{2\log(K/\alpha)}W_k|D_k|\right\} \equiv V_K.
\end{aligned} \tag{9}$$

Since $F_k(p) = E[h(D_K, W_K, p)]$, it follows from (7) and the above inequalities that

$$\begin{aligned}\sup_{\alpha/K \le p \le \alpha} \left| \frac{F_k(p)}{p} - 1 \right| &= \sup_{\alpha/K \le p \le \alpha} \left| E\left(\frac{h(D_k, W_k, p)}{p} - 1 \right) \right| \\ &\le \sup_{\alpha/K \le p \le \alpha} E \left| \frac{h(D_k, W_k, p)}{p} - 1 \right| \\ &\le E(U_K) + E(V_K).\end{aligned}$$

Now $U_k \xrightarrow{p} 0$ and $V_k \xrightarrow{p} 0$ since $D_k = o_p(1/\log K) = |W_k - 1|$ (from the $\sqrt{K}$-consistency of the U-statistics). Lemma 3 shows that U_K and V_K are uniformly integrable. It follows that $E(U_K) \to 0$ and $E(V_K) \to 0$. $\diamond$

PROOF OF THEOREM 2. It follows from Benjamini and Yekutieli (2001), that, for each $k \in S$, there exists a partition $\{C_r^k : r = 1, \ldots, K\}$ of the sample space such that

$$E(Q) = \sum_{k \in S_0} \sum_{r=1}^{K} \frac{1}{r} Pr\left(\hat{P}_k \le \frac{r}{K}\alpha \right) Pr(C_r^k). \tag{10}$$

So,

$$\begin{aligned}E(Q) &= \sum_{k \in S_0} \sum_{r=1}^{K} \frac{1}{r} Pr\left(\hat{P}_k \le \frac{r}{K}\alpha \right) Pr(C_r^k) \\ &= \sum_{k \in S_0} \sum_{r=1}^{K} \frac{1}{r} F_k\left(\frac{r}{K}\alpha \right) Pr(C_r^k) \\ &= \sum_{k \in S_0} \sum_{r=1}^{K} \frac{1}{r} \frac{F_k\left(\frac{r\alpha}{K}\right)}{\frac{r\alpha}{K}} \frac{r\alpha}{K} Pr(C_r^k) \\ &\le \frac{\alpha}{K} \sup_{\alpha/K \le p \le \alpha} \frac{F_k(p)}{p} \sum_{k \in S_0} \sum_{r=1}^{K} Pr(C_r^k) \\ &\le \alpha \sup_{\alpha/K \le p \le \alpha} \frac{F_k(p)}{p}. \diamond\end{aligned}$$

PROOF OF THEOREM 3. This follows from Theorem 2 and Lemma 2. $\diamond$

The following result is needed for Lemma 2.

LEMMA 3. The quantities U_K and V_K defined in (8) and (9) are uniformly integrable.

PROOF. It suffices to show that $\limsup_K E(U_K^2) < \infty$ and $\limsup_K E(V_K^2) < \infty$. We will show this for the non-robust version of the estimators $\hat{\tau}$ and $\hat{\mu}$. Recall that

$$U_K = C\alpha^{W_k^2 - 1} |D_k| K^{1 - W_k^2} \left(\log(K/\alpha)\right)^{W_k^2/2} \exp\left\{ 2\sqrt{2\log(K/\alpha)} W_k |D_k| \right\}.$$

Now,

$$W_k = \frac{\hat{\tau}_k}{\tau} = 1 + \frac{\hat{\tau} - \tau}{\tau}.$$

By the law of large numbers for U-statistics, $W_k \stackrel{a.s}{\to} 1$. By the law of the iterated logarithm for U-statistics,

$$\limsup_K \sqrt{\frac{K}{2 \log \log K}} \, (\hat{\tau} - \tau) < C \;\; a.s.$$

for some finite C. (Here, C will denote a generic, positive constant that need not be the same in each expression.) Hence, for all large K,

$$W_k \leq 1 + 2C \sqrt{\frac{2 \log \log K}{K}}$$

almost surely. Thus, eventually,

$$(K^{1-W_k^2})^2 = K^{2(1-W_k)(1+W_k)} \leq \exp \left\{ C \log K \sqrt{\frac{2 \log \log K}{K}} \right\} \to 0.$$

The rest of the factors involving W_k can be bounded similarly. Similarly, $|D_k| \leq C\sqrt{\frac{2 \log \log K}{K}}$ for all large k almost surely and these factors can thus also be bounded. Hence, $\limsup_K E(U_K^2) < \infty$ as required. The argument is the same for V_K.

Acknowledgment

Jung-Ying Tzeng is a PhD candidate, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. W. Byerley is Professor, Department of Psychiatry, University of California, Irvine, CA. B. Devlin is Assistant Professor, Department of Psychiatry, University of Pittsburgh, PA. Kathryn Roeder is Professor of Statistics, Carnegie Mellon University. Larry Wasserman is Professor of Statistics, Carnegie Mellon University. This research was supported by National Institute of Health Grant MH57881 and National Science Foundation Grant DMS-0104016.

Reference

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B* 57, 289-300.

Benjamini, Y., Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," Ann Stat 29, 1165-1188.

Devlin, B. and Roeder, K. (1999), "Genomic control for association studies," *Biometrics* 55, 997-1004.

Devlin, B., Roeder, K., Otto, C., Tiobech, S. and Byerley, W. (2001), "Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania," *Human Genetics*.

Devlin, B., Roeder, K. and Wasserman, L. (2000) Genomic Control for Association Studies: A semiparametric test to detect excess-haplotype sharing. *Biostatistics* 1:369-387.

Devlin B, Roeder K, Wasserman L. 2001. Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology 60:156-166.

Genovese, G. and Wasserman, L. (2001), "Operating Characteristics and Extensions of the FDR Procedure," Department of Statistics Technical Report # 737, Carnegie Mellon University.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure." Scandinavian Journal of Statistics, 6, 65-70.

McPeek, M.S. and Strahs, A. (1999), "Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping," *Amer. J. Human Genet.* 65, 858-875.

Ott, J.(2000), "Predicting the range of linkage disequilibrium," *Proc Natl Acad Sci U S A* 97, 2-3.

Risch N, and Merikangas K. (1996), "The future of genetic studies of complex human diseases," *Science* 273, 1516-1517.

Simmons, R., Graydon, J., Gajdusek, D., and Brown, P. (1965), “Blood group genetic variations in natives of the Caroline Islands and in other parts of Micronesia,” *Oceania* 36, 132-170.

Slager S.L., and Schaid D.J. (2001), “Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects,” *Am J Hum Genet* 68, 1457-1462.

Sobel, E. and Lange, K. (1996), “Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics,” *Am. J. Hum. Genet.* 58, 1323-1337.

Takayama, J. (1981), “Early pottery and population movements in Micronesian prehistory,” *Asian Perspectives* 24, 1-10.

	5% outliers				20% outliers			
q	1	2	3	4	1	2	3	4
μ_k	0.015	0.025	0.006	0.012	0.015	0.025	0.006	0.012
$\hat{\mu}$	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001
$\hat{\tau}$	1.078	1.079	1.053	1.083	1.643	1.631	1.403	1.612
FDR	0.007	0.008	0.009	0.007	0.000	0.000	0.000	0.000
power	0.985	0.999	0.396	0.888	0.917	0.984	0.133	0.539

Table 1. Simulation Results of Applying Matching Statistic and FDR Procedure.

	q	Matching Stat	Chi-square Stat
Type I Error			
	1	0.000	0.001
	2	0.000	0.001
	3	0.000	0.001
	4	0.000	0.001
Power			
	1	0.965	0.997
	2	0.997	0.880
	3	0.354	0.948
	4	0.821	0.602

Table 2. Comparison of the Matching Statistic and the Omnibus Chi-square Test Statistic Using Holm's Correction.

ID	M1	M2	M3	M4	M5	M6	M7	M8
CASES								
1	3	0*	7	2	6	11	3	1
1	3	0	6	3	5	10	3	6
2	2	0	5	4	3	10	3	5
2	2	0	4	3	0	10	4	3
				.				
				.				
				.				
22	3	8	3	3	4	14	1	3
22	3	8	4	3	3	14	4	2
CONTROLS								
1	3	0	6	3	5	10	3	6
1	5	0	7	1	3	14	2	7
				.				
				.				
				.				
27	2	0	5	7	5	14	2	7
27	5	8	7	2	3	11	3	5

*: “0” denotes missing data

Table 3. Data matrix: shown by Chromosome 22 as an example. Each row represents a haplotype and the pairs of haplotypes reveal the complete set of STR alleles for an individual. The markers are STRs and the numbers represent the number of repeats of particular STR markers at each of 8 loci.

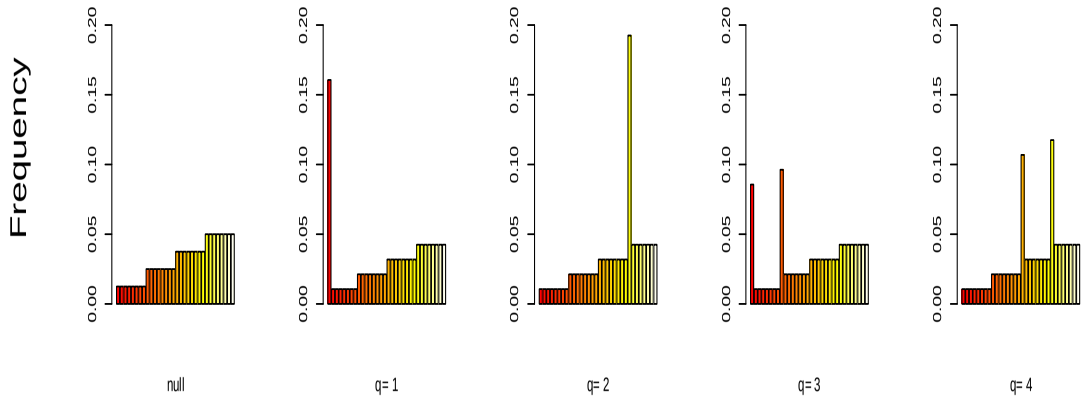
Figure Captions

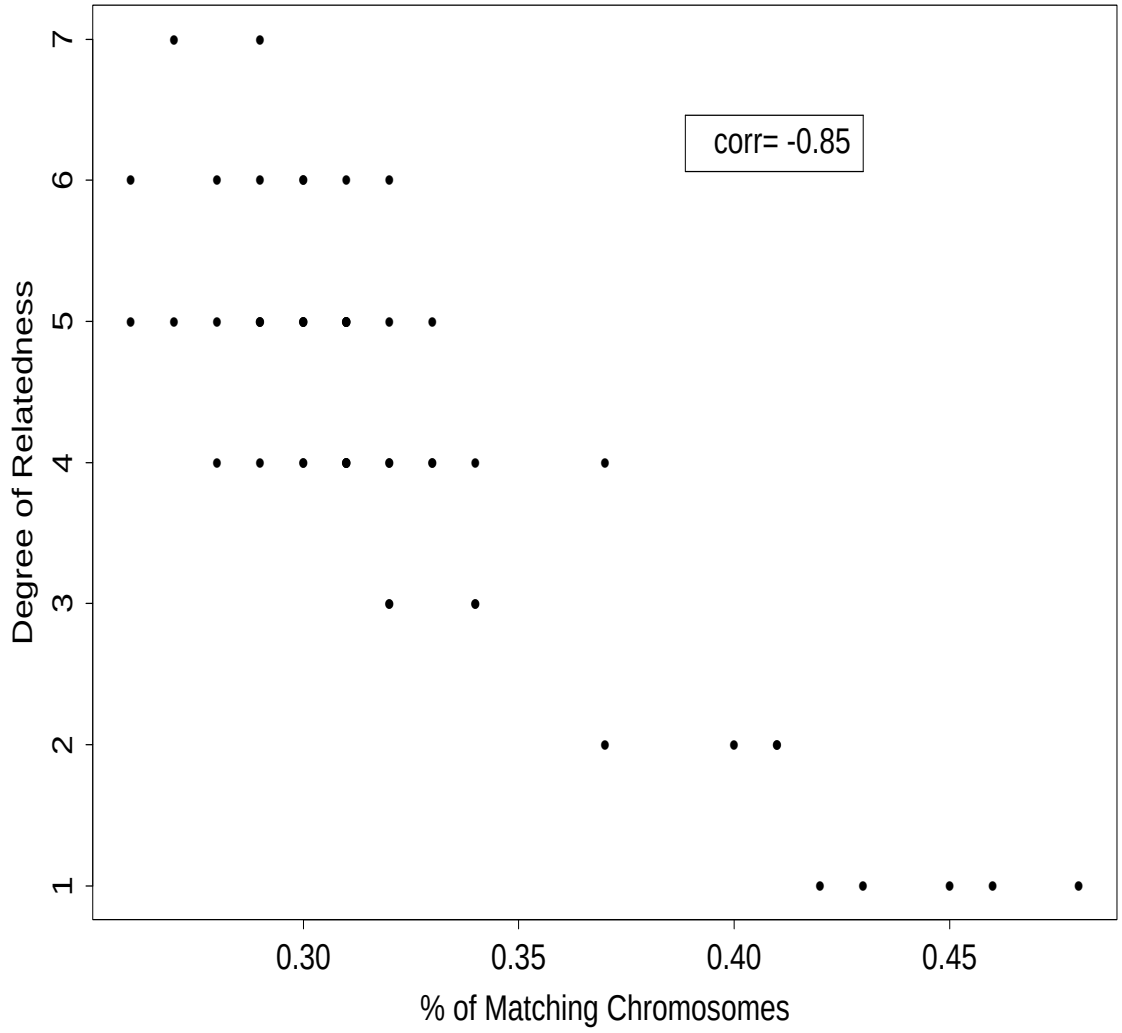
Figure 1. Haplotype distribution for simulations. From left to right, under the null and four alternative hypotheses.

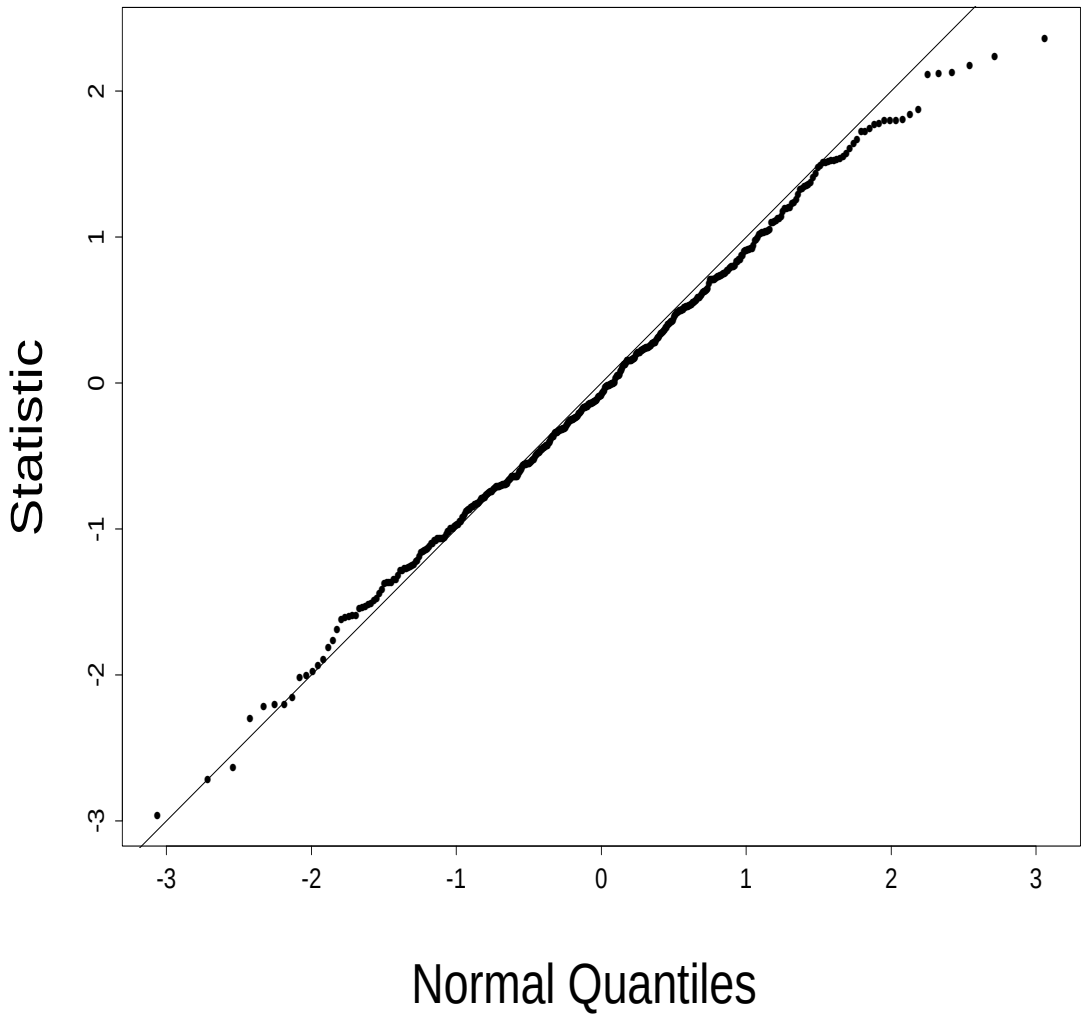
Figure 2. Relationship between the degree of relatedness and the degree of haplotype similarity.

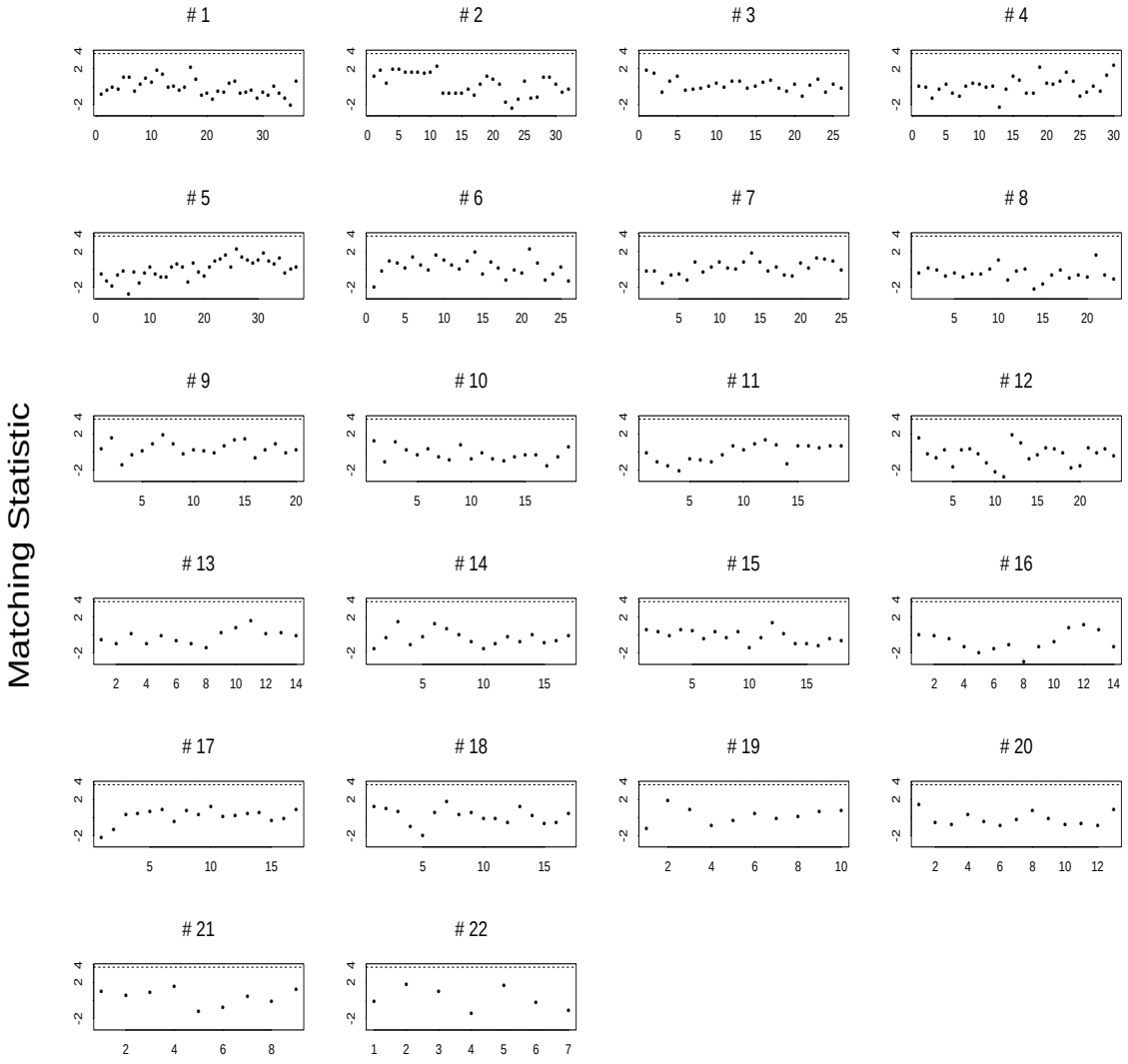
Figure 3. Normal scores plot for the matching statistic.

Figure 4. Matching statistic plotted as a function of the relative haplotype locations on the 22 autosomal chromosomes. Markers are roughly located on a 10cM grid; the numbers on the abscissa indicate a moving window of marker pairs. For chromosome 22, which has 8 loci, 7 statistics are given for 2-locus haplotypes formed from loci 1-2, 2-3, ..., 7-8. The autosomal chromosomes 1-22 are depicted from left to right. The dashed-lines denote the significant level using a Bonferroni correction with α=.05.


Frequency
0.00
0.05
0.10
0.15
0.20
null
q= 1
q= 2
q= 3
q= 4


corr= -0.85
Degree of Relatedness
% of Matching Chromosomes
1
2
3
4
5
6
7
0.30
0.35
0.40
0.45


Statistic
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Normal Quantiles


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Matching Statistic
Region