Statistics
in Medicine

Research Article

Received 2 March 2011, Accepted 10 August 2011 Published online 23 September 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4397

Pseudosibship methods in the case-parents design

Zhaoxia Yu[a*†] **and Li Deng**[b]

Recent evidence suggests that complex traits are likely determined by multiple loci, each of which contributes a weak to moderate individual effect. Although extensive literature exists on multilocus analysis of unrelated subjects, there are relatively fewer strategies for jointly analyzing multiple loci using family data. Here we address this issue by evaluating two pseudosibship methods: the 1:1 matching, which matches each affected offspring to the pseudosibling formed by the alleles not transmitted to the affected offspring, and the exhaustive matching, which matches each affected offspring to the pseudosiblings formed by all the other possible combinations of parental alleles. We prove that the two matching strategies use exactly and approximately the same amount of information from data under additive and multiplicative genetic models, respectively. Using numerical calculations under a variety of models and testing assumptions, we show that compared with the exhaustive matching, the 1:1 matching has comparable asymptotic power in detecting multiplicative/additive effects in single-locus analysis and main effects in multilocus analysis, and it allows association testing of multiple linked loci. These results pave the way for many existing multilocus analysis methods developed for the case-control (or matched case-control) design to be applied to case-parents data with minor modifications. As an example, with the 1:1 matching, we applied an L_1 regularized regression to a Crohn's disease dataset. Using the multiple loci selected in our approach, we obtained an order-of-magnitude decrease in p-value and an 18.9% increase in prediction accuracy when compared with using the most significant individual locus.

Keywords: case-parents; pseudocontrols/siblings; transmission/disequilibrium; multilocus

1. Introduction

It has long been recognized that spurious marker disease association can be introduced by population stratification in case-control studies [1]. When subjects are sampled from multiple populations that are different in both disease prevalence and allele frequencies, association analysis based on a case-control design might lead to false findings. As a result, population stratification is one of the most often cited explanations for the difficulties in replicating results of genetic association studies [2–4]. One way to avoid drawing erroneous conclusions on genetic associations is to use family-based designs, such as the case-parents design, in which both cases and their parents are genotyped [5]. The validity of this design stems from the fact that the hypothetical control formed by the nontransmitted alleles in a case-parents trio is from the same population as the case is.

When testing the association between a locus and a disease using case-parents data, modeling the genotype of an offspring conditioning on parental mating (genotypes) and the offspring being affected prevents spurious associations caused by population stratification [6, 7]. The conditional approach is equivalent to matching the affected offspring to his or her three pseudosiblings formed by the other three combinations of parental alleles (hereafter 1:3 matching), and conditional logistic regressions can be used for statistical inference [6, 7]. Different tests based on the conditional logistic regression framework have been proposed, such as likelihood ratio tests (LRT) and score tests under a variety of assumptions about the true genetic model [6, 8, 9]. It has been shown that the well-known transmission disequilibrium test (TDT) [10, 11] is the score test of the conditional logistic regression using the 1:3 matching under multiplicative model [8]. Interestingly, as originally derived, the TDT is also a McNemar [12] type of

[a]*Department of Statistics, University of California, Irvine, CA 92697, USA*
[b]*Department of Vision Science, New England College of Optometry, Boston, MA 02115, USA*
**Correspondence to: Zhaoxia Yu, Department of Statistics, University of California, Irvine, CA 92697, USA.*
[†]*E-mail: zhaoxia@ics.uci.edu*

test when matching each affected offspring to his or her pseudosibling formed by the alleles that are not transmitted to the affected offspring (hereafter 1:1 matching). Other tests based on the 1:1 matching include the tests introduced by Terwilliger and Ott [11], and Wittkowski and Liu [13]. Obviously, using three pseudosiblings for each case utilizes more information from data than the 1:1 matching. Cordell and Clayton [14] found that the 1:3 matching is more efficient than the 1:1 matching.

The 1:3 matching in a single-locus analysis is exhaustive in that each affected offspring is matched to all the other possible combinations of parental alleles. Generalized to L unlinked loci, the number of all pseudosiblings matched to an affected child then is $4^L - 1$ [14, 15]. In computation, this exponential growth can limit the number of loci to be jointly analyzed. For linked loci, additional computational complexity arises, as the transmission of parental alleles to offspring does not follow Mendel's law of independent assortment, and all the possible offspring genotypes of a pair of parents do not occur with equal probabilities. Therefore, the recombination fractions between SNPs are required to build conditional logistic regressions [14–16]. However, these recombination fractions are usually unknown and difficult to estimate. Thus, although the exhaustive matching maximizes information extracted from the case-parents data, it is difficult to use in multilocus association analysis. To avoid those complications, instead of the exhaustive matching, we can use the 1:1 matching.

In the literature, a number of multilocus methods using the 1:1 matching have been proposed [17–30]. In those methods, a test statistic is first computed by comparing the transmitted alleles to the nontransmitted alleles, and then the statistical significance is assessed either by asymptotic theories or by a permutation procedure that randomly shuffles the 'transmitted' and 'nontransmitted' labels of each affected offspring and his or her pseudosibling. In terms of haplotype phase, both haplotype-based [17–21, 23, 24] and genotype-based [22, 24–30] methods have been proposed. For genotype-based methods, both main effects [22, 24, 27] and gene–gene interaction have been tested [26]. Although the 1:1 matching is more straightforward to implement, computationally more tractable, and allows association testing of multiple linked single nucleotide polymorphisms (SNPs), there is no doubt that it utilizes less information from data than the exhaustive matching. It is therefore important to know the efficiency of the 1:1 matching relative to that of the exhaustive matching. Especially, if the efficiency of the 1:1 matching is model-dependent or test-dependent, it is critical to evaluate under what situations the 1:1 matching is efficient.

In this paper, we systematically investigate the relative efficiency of the two ways of creating pseudocontrols: the 1:1 matching and the exhaustive matching. We prove that the two matching strategies use exactly and approximately the same amount of information from data under additive and multiplicative models, respectively. We also quantify the efficiency of the two matching strategies using statistical power computed from asymptotic LRTs under the conditional logistic regression framework. We compare the efficiency of the two matching strategies under different genetic models and testing models, in both one-locus analysis and two-locus analysis. As an illustration of using the 1:1 matching to conduct multilocus analysis, we apply an L_1 regularized variable selection method to a Crohn's disease dataset and compare its performance to using the most significant single locus.

2. Methods

2.1. The equivalence of the 1:1 matching and the exhaustive matching under additive genotype relative risks

Suppose we are testing the association between a disease and an SNP with a risk allele 'A' and a normal allele 'a'. For convenience, we use numerically coded genotypes based on the number of the 'A' alleles and adopt the notations of trio types used in [9]. If we ignore the order of parents, there are six parental mating types and 10 trio types, which are illustrated in the first two columns of Table I. For a case-parents trio, let $t \in \{12, 22, 21, 31, 42, 41, 40, 51, 50, 60\}$ denote the trio type determined by the parental mating type and the genotype of the affected offspring, as shown in Table I. When testing the association between a locus and a disease using the case-parents design, the approach conditioning on parental mating type and the case being affected is equivalent to the 1:3 matching, which matches each affected offspring to his or her three pseudosiblings [6, 7]. We use $H_E(t)$ to denote the 1:3 matching set of numerically coded genotypes that comprises the genotype of the offspring and the other three combinations of parental alleles. Generalized to L unlinked SNPs, the set $H_E(.)$ consists of 4^L genotypes. An alternative matching strategy matches each affected offspring to his or her pseudosibling constructed by the alleles that are not transmitted to the affected offspring and we denote the 1:1 matching set by

Table I. Trio type, parental mating type, case type, and pseudocontrols for case-parents data. The first column shows the trio types, which are determined by parental mating type and case genotype. The second column shows the parental mating types, and if we ignore the order of parents, there are six types in total, as shown from top to bottom. The third column shows the genotypes of the affected offspring; there are three types in total, AA, Aa and aa, which are represented by type 2, 1, and 0, respectively. The fourth column shows the genotypes of the pseudosiblings of the affected offspring under the 1:3 matching; the genotype in bold letters is that of the pseudosibling of the affected offspring under the 1:1 matching.

Trio type	Parental mating type	Case type	Pseudocontrols
12	1: AA × AA	2: AA	**AA**, AA, AA
22	2: AA × Aa	2: AA	**Aa**, AA, Aa
21		1: Aa	**AA**, AA, Aa
31	3: AA × aa	1: Aa	**Aa**, Aa, Aa
42	4: Aa × Aa	2: AA	**aa**, Aa, Aa
41		1: Aa	**Aa**, AA, aa
40		0: aa	**AA**, Aa, Aa
51	5: Aa × aa	1: Aa	**aa**, Aa, aa
50		0: aa	**Aa**, Aa, aa
60	6: aa × aa	0: aa	**aa**, aa, aa

$H_1(t)$. As an example, suppose the genotypes of the affected offspring, the father, and the mother are 2 (AA), 1 (Aa), and 1 (Aa), respectively, then $t = 42$, $H_1(t) = \{2, 0\}$ and $H_E(t) = \{2, 0, 1, 1\}$, as shown in Figure 1.

Let g_l, $l = 1, 2, \ldots, L$ be the number of copies of the risk allele at SNP l. We define the genotype relative risk (GRR) of genotype $(g_1, g_2, \ldots, g_L)$ as the relative risk of being affected for subjects with this genotype relative to subjects with genotype $(0, 0, \ldots, 0)$:

$$r_{g_1,g_2,\ldots,g_L} = \frac{\Pr(D = 1|g_1, g_2, \ldots, g_L)}{\Pr(D = 1|0, 0, \ldots, 0)}$$

Definition 1
We say the GRRs are additive if

$$r_{g_1,g_2,\ldots,g_L} = 1 + \sum_{l=1}^{L} \gamma_l g_l$$

where $\gamma_l = r_{0,0,\ldots,1,\ldots,0} - 1 = \frac{\Pr(D=1|0,0,\ldots,g_l=1,\ldots,0)-\Pr(D=1|0,0,\ldots,0)}{\Pr(D=1|0,0,\ldots,0)}$, for $l = 1, \ldots, L$. For example, when $L = 1$, $r_0 = 1, r_1 = 1 + \gamma_1, r_2 = 1 + 2\gamma_1$.

Note that additive GRRs imply additive penetrances. We next show that additive GRRs imply the equivalence between the 1:1 matching and the exhaustive matching.

Theorem 1
Suppose the GRRs are additive, the 1:1 matching and the exhaustive matching use exactly the same amount of information from data. In other words, the likelihood functions of the 1:1 matching and the exhaustive matching are identical.

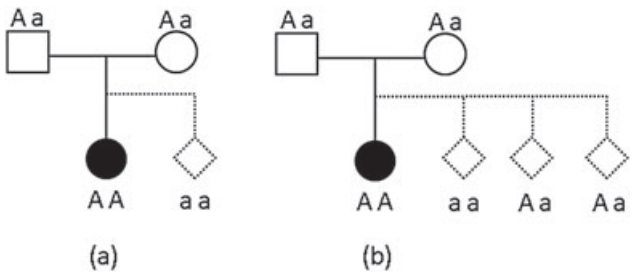

Figure 1. One (a) versus three (b) pseudosiblings for the affected offspring with AA genotype, Aa father, and Aa mother. The filled circles represent the genotypes of the affected offspring, and the dotted diamonds represent the genotypes of the pseudosiblings.

The proof for unlinked loci is provided in Appendix A. When loci are linked, if we also know recombination fractions among loci, we can use a weighted exhaustive matching, with the weights determined by recombination fractions. With the same method used in Appendix A and some additional notations, we can prove that the 1:1 matching and the weighted exhaustive matching use the same amount of information from data under additive GRRs. Here we focus only on unlinked loci for two reasons. First, this situation leads to the largest difference between the two matching strategies. Second, when SNPs are linked, using the exhaustive matching requires recombination fractions [14–16] that might not be easily estimated from case-parents data.

To make the results intuitive, consider the situation when L is 1. Let n_t denote the number of case-parents trios in trio type t, where $t \in \{12, 22, 21, 31, 42, 41, 40, 51, 50, 60\}$. The likelihood functions under the exhaustive matching and under the 1:1 matching are

$$L_E(r_1, r_2) \propto \left(\frac{r_2}{r_1+r_2}\right)^{n_{22}} \left(\frac{r_1}{r_1+r_2}\right)^{n_{21}} \left(\frac{r_2}{1+2r_1+r_2}\right)^{n_{42}} \left(\frac{2r_1}{1+2r_1+r_2}\right)^{n_{41}} \left(\frac{1}{1+2r_1+r_2}\right)^{n_{40}} \left(\frac{r_1}{1+r_1}\right)^{n_{51}} \left(\frac{1}{1+r_1}\right)^{n_{50}} \tag{1}$$

and

$$L_1(r_1, r_2) \propto \left(\frac{r_2}{r_1+r_2}\right)^{n_{22}} \left(\frac{r_1}{r_1+r_2}\right)^{n_{21}} \left(\frac{r_2}{1+r_2}\right)^{n_{42}} \left(\frac{1}{1+r_2}\right)^{n_{40}} \left(\frac{r_1}{1+r_1}\right)^{n_{51}} \left(\frac{1}{1+r_1}\right)^{n_{50}} \tag{2}$$

respectively. Under the assumption of additive GRRs, that is, $1 + r_2 = 2r_1$, both L_E and L_1 are proportional to

$$\left(\frac{2r_1-1}{3r_1-1}\right)^{n_{22}} \left(\frac{r_1}{3r_1-1}\right)^{n_{21}} \left(\frac{2r_1-1}{2r_1}\right)^{n_{42}} \left(\frac{1}{2r_1}\right)^{n_{40}} \left(\frac{r_1}{1+r_1}\right)^{n_{51}} \left(\frac{1}{1+r_1}\right)^{n_{50}}$$

Definition 2
We say GRRs are multiplicative if

$$r_{g_1,g_2,\ldots,g_L} = \prod_{l=1}^{L} (1+\gamma_l)^{g_l}$$

Corollary 1
When the marginal genetic effects are not large, the 1:1 matching and the exhaustive matching use approximately the same amount of information from data under multiplicative GRRs.

When the marginal genetic effects are not large, we can use the first-order Taylor expansion to approximate the GRRs and get rid of terms of higher orders. Hence, the result follows immediately. Therefore, the 1:1 matching and the exhaustive matching also use similar amount of information from data under multiplicative GRRs. When the frequency of the risk allele at a locus is small, the dominant model is close to the multiplicative model, implying the near equivalence between the two matching strategies under the dominant model. It is not clear how much power the 1:1 matching would lose under the recessive model. To quantify the relative efficiency of the 1:1 matching to that of the exhaustive matching, we use numerical comparisons under a variety of models and testing assumptions for both one-locus and two-locus analysis, as described next.

2.2. *One-locus models and tests*

2.2.1. Likelihood ratio tests for the association between a disease and an SNP. For testing the association between an SNP and a disease, the conditional likelihood function of conditioning on parental mating type and the case being affected is equivalent to the 1:3 matching likelihood function $L_E(r_1, r_2)$, given in (1). An alternative matching method is the 1:1 matching, with the likelihood function denoted by $L_1(r_1, r_2)$, provided in (2). On the basis of the likelihood functions, we consider the following three types of likelihood ratio tests.

The first type of tests we study is two-degree-of-freedom tests. In this type of tests, we compare the maximized likelihood with parameters r_1 and r_2 to the likelihood under the null hypothesis $r_1 = r_2 = 1$.

We denote the resulting 2-DOF test with the 1:3 matching as '$2df_E$' and the test with the 1:1 matching as '$2df_1$'.

The second type of tests we study has one DOF, which is obtained by placing restrictions on GRRs. We consider the following three commonly used restrictions:

- Multiplicative, for which $r_2 = r_1^2$
- Recessive, for which $r_1 = 1$
- Dominant, for which $r_2 = r_1$

The likelihood ratio tests are constructed by comparing the maximized likelihood under each restriction to the likelihood under the null hypothesis $r_1 = r_2 = 1$. In each of the above restricted models, only one parameter needs to be estimated, therefore the resulting likelihood ratio tests have only one DOF. We denote the three tests with the 1:3 matching as 'MUL_E', 'REC_E', and 'DOM_E', respectively; and denote the three tests with the 1:1 matching as 'MUL_1', 'REC_1', and 'DOM_1', respectively. Because the TDT [10] is widely used, we also compute the power of the TDT.

The third type of tests we consider is the Hardy–Weinberg equilibrium (HWE) test. Testing HWE in case-only data or case-control data has been used in fine-scale mapping [31–33]. It has also been used to provide information about the underlying disease models and thus to improve statistical power [34–36]. In case-only data or case-control data, testing HWE is equivalent to testing whether the underlying genotype penetrance is multiplicative because HWE in affected subjects holds if and only if the disease model is multiplicative. In case-parents data, multiplicative GRRs imply independent transmission of alleles from parents to the affected offspring [9], that is, the transmission follows HWE. Therefore, we can conduct the HWE test for case-parents data by testing whether the GRRs are multiplicative. Under the null hypothesis of multiplicative GRRs, the likelihood functions using the 1:3 matching and the 1:1 matching are

$$L_E(r) \propto \left(\frac{r}{1+r}\right)^{n_{22}} \left(\frac{1}{1+r}\right)^{n_{21}} \left(\frac{r^2}{(1+r)^2}\right)^{n_{42}} \left(\frac{2r}{(1+r)^2}\right)^{n_{41}} \left(\frac{1}{(1+r)^2}\right)^{n_{40}} \left(\frac{r}{1+r}\right)^{n_{51}} \times \left(\frac{1}{1+r}\right)^{n_{50}}$$

$$\text{and } L_1(r) \propto \left(\frac{r}{1+r}\right)^{n_{22}} \left(\frac{1}{1+r}\right)^{n_{21}} \left(\frac{r^2}{1+r^2}\right)^{n_{42}} \left(\frac{1}{1+r^2}\right)^{n_{40}} \times \left(\frac{r}{1+r}\right)^{n_{51}} \left(\frac{1}{1+r}\right)^{n_{50}}$$

respectively. For the 1:3 matching, we construct the likelihood ratio test by comparing the maximized $L_E(r_1, r_2)$ to the maximized $L_E(r)$, and we denoted the resulting test as 'HWE_E'. Similarly, for 1:1 matching, we construct the likelihood ratio test 'HWE_1' by comparing the maximized $L_1(r_1, r_2)$ to the maximized $L_1(r)$.

2.2.2. Power calculation. We have introduced nine likelihood ratio tests for both the 1:3 matching and the 1:1 matching. The efficiency depends not only on tests, but also on true genetic models. Here, we consider four true genetic models: multiplicative, recessive, dominant, and additive. The definitions of the models have been discussed previously.

To compute the asymptotic power, we choose the number of trios to be 200 and fix the frequency of the risk allele 'A' to 0.2. We assume that the trios are from a random mating population. For a given true genetic model, we first compute the expected numbers of different trio types. Using the expected numbers of trio types, for a given test, we then compute the noncentrality parameter using twice the log of the ratio of the likelihood maximized under the alternative hypothesis to the likelihood maximized under the null hypothesis [15]. As an example, to compute the power of the $2df_E$ test under a multiplicative model, we first compute the expected numbers of trio types under the multiplicative model then compute the noncentrality parameter $ncp = 2\log(\max L_E(r_1, r_2)/L_E(1,1))$, where $\max L_E(r_1, r_2)$ is the likelihood maximized over the parameters r_1 and r_2. Finally, we compute the power using $\Pr(X_2(ncp) > \chi^2_{2,0.95})$, where $X_2(ncp)$ is a chi-square-distributed random variable with two DOFs and the noncentrality parameter ncp and $\chi^2_{2,0.95}$ is the 95 th percentile of the chi-square distribution with two DOFs.

2.3. *Two-locus models*

We use two-locus models as examples of multilocus scenarios because multilocus models with more than two loci will result in complicated exhaustive matching. In the two-locus analysis, we consider both the 1:3 matching and the 1:15 (exhaustive) matching. We assume that the two SNPs are unlinked because this situation leads to the largest difference between the two matching strategies.

Two SNPs can jointly affect the risk of a disease in many different ways. Here, we consider four types of true genetic models:

mul–mul: the model with multiplicative main effects at both SNPs and no interaction;
dom–dom: the model with dominant main effects at both SNPs and no interaction;
rec–rec: the model with recessive main effects at both SNPs and no interaction; and
gene–gene interaction: the model with multiplicative main effects and gene–gene interaction effect.

The GRRs of the four models are shown in Table II.

To test the genetic association between the two loci and a disease, we consider the following conditional likelihood function

$$L_H(\alpha,\beta,\gamma)=\prod_{i=1}^{n}\frac{\exp(\alpha g_{o,i}+\beta h_{o,i}+\gamma g_{o,i}h_{c,i})}{\sum_{(g'_{o,i},h'_{o,i})\in \mathrm{H}(t_{i1},t_{i2})}\exp(\alpha g'_{o,i}+\beta h'_{o,i}+\gamma g'_{o,i}h'_{o,i})}$$

where H represents the 1:1 matching or the exhaustive matching; α, β, and γ denote the coefficients of the main effect at the first SNP, the main effect at the second SNP, and the interaction effect between the two SNPs, respectively; $g_{o,i}$ and $h_{o,i}$ are the numerically coded genotypes of the affected offspring in the i-th case-parents trio at the first and the second SNPs, respectively; and t_{i1} and t_{i2} are the trio types defined based upon the first and the second SNPs, respectively.

For the first three genetic models, namely mul–mul, dom–dom, and rec–rec, no interaction is assumed between the two loci. We test the main effect using LRTs similar to those described in the one-locus analysis. Specifically, we compare $L_H(\alpha,\beta,0)$ at its maximized value to $L_H(0,0,0)$. For the gene–gene interaction model, in addition to testing the main effect, we also test the gene–gene interaction by comparing the maximized value of $L_H(\alpha,\beta,\gamma)$ to that of $L_H(\alpha,\beta,0)$. The computation of power of the LRTs under the four true genetic models are similar to what we use in the one-locus analysis, with 200 case-parents trios sampled from a random mating population with the risk allele frequencies at both SNPs being 0.2.

Table II. The GRRs of four two-locus genetic models: mul–mul, dom–dom, rec–rec, and gene–gene interaction model, from top to bottom, respectively.

		SNP2		
Genetic model	SNP1	bb	Bb	BB
	aa	1	λ	λ^2
mul–mul	Aa	λ	λ^2	λ^3
	AA	λ^2	λ^3	λ^4
	aa	1	λ	λ
dom–dom	Aa	λ	λ^2	λ^2
	AA	λ	λ^2	λ^2
	aa	1	1	λ
rec–rec	Aa	1	1	λ
	AA	λ	λ	λ^2
gene–gene	aa	1	λ	λ^2
interaction:	Aa	λ	λ^2	$1.2\lambda^3$
Test for main effects	AA	λ^2	$1.2\lambda^3$	$1.2^2\lambda^4$
gene–gene	aa	1	1.2	1.2^2
interaction:	Aa	1.2	$1.2^2\lambda$	$1.2^3\lambda^2$
Test for interaction	AA	1.2^2	$1.2^3\lambda^2$	$1.2^4\lambda^4$

2.4. Application to real data

We conduct both single-locus and multilocus analyses on a Caucasian-based case-parents dataset in which all offspring had Crohn's disease. Crohn's disease is an inflammatory bowel disease that affects about half a million people in North America [37]. Using a linkage analysis of nuclear families with inflammatory bowel disease patients, Rioux *et al.* [38] detected linkage signals at human chromosome 5q31. To narrow down the candidate region, they genotyped SNPs at 5q31 for 139 case-parents trios [39]. We used the publically available subset, that is, 129 trios genotyped at 103 common SNPs, available from http://www.broadinstitute.org/archive/humgen/IBD5/haplodata.html (data downloaded on 27 July 2010). These 103 SNPs cover a 500-kb region at 5q31 and the linkage disequilibrium structure of the 103 SNPs was reported previously [40]. In our analysis we excluded two trios with more than 40% missing genotypes and analyzed the remaining 127 trios.

For the single-locus analysis of the data, we consider four types of tests, with each assuming a specific genetic model: 2*df*, the 2-DOF test; *DOM*, the dominant test; *MUL*, the multiplicative test; and *REC*, the recessive test. Each type is conducted with both the 1:1 matching and the 1:3 matching, so in total there are eight tests. All the eight tests are applied to each of the 103 SNPs of the Crohn's disease dataset.

We also conduct the multilocus analysis using the 1:1 matching. The exhaustive matching cannot be used, because the recombination fractions are unknown and cannot be accurately estimated, and there are 103 SNPs. In the Crohn's disease data, the number of SNPs is close to the number of trios, which would create problems if we jointly analyze all the SNPs in a conditional logistic regression. This motivates us to use the Lasso regularization [41]. Another advantage of the Lasso regularization is that it shrinks estimates of regression coefficients toward zero by adding an L_1 penalty term; therefore, it is especially suitable for selecting SNPs in genetic association analysis [42–45]. To realize Lasso regularization, we use the *penalized* R (www.r-project.org) package, which provides L_1 regularized estimation for Cox proportional-hazards regression [46]. The reason why we chose this package is that Cox proportional-hazards regression can be used to fit conditional logistic regression by putting each matched case-control pair in a unique stratum [9]. In our analysis, we first impute missing genotype data using BEAGLE [47] because the *penalized* package requires complete data. Then with the genotype variable of each SNP being numerically coded to be 0, 1, or 2 according to the number of copies of the rare allele of the SNP, we use the leave-one-out cross-validation to choose the optimal tuning constant, which determines the degree of shrinkage of the regression coefficients. Following that, we compute the coefficients of the conditional logistic regression for the SNPs chosen with the optimal tuning constant. And finally, we compute the percentage of correctly predicted disease status. For comparison, we also compute the percentage of correctly predicted disease status using the most significant SNP.

3. Results

3.1. Results of one-locus models

Figures 2 and 3 show the power of the TDT and the four tests under four different one-locus genetic models for both the 1:1 matching and the 1:3 matching. The power of the TDT and MUL_E was nearly identical, and their power curves almost overlap each other in Figures 2 and 3. This is not surprising because the TDT is the score test and MUL_E is the likelihood ratio test of the same model. We also subtract the power of the 1:1 matching from that of the 1:3 matching and summarize the differences in Table III. The positive values in Table III indicate that the 1:3 matching is more powerful than the 1:1 matching, whereas the negative values indicate the opposite. When the testing model agrees with the true model, because the 1:3 matching uses more information from data, naturally, the 1:3 matching is always more powerful than the 1:1 matching. It is also more efficient under the 2-DOF test for all true genetic models except the additive model. This exception is due to the fact that the likelihood functions $L_E(r_1, r_2)$ and $L_1(r_1, r_2)$ are identical under additive GRRs.

There are also many situations wherein the power of the 1:1 matching is comparable to or even greater than that of the 1:3 matching. For example, when the true model is additive, the two matching approaches have the same power under the 2-DOF test, and the 1:1 matching is more powerful under the other tests. When the true model is multiplicative, the 1:1 matching is also more powerful under the dominant or the recessive test, and it is only slightly less powerful under the multiplicative or the 2-DOF test. Another situation where the 1:1 matching is superior is when the testing model is quite different from the true model. For example, when the testing model is dominant but the true model is recessive, as shown in

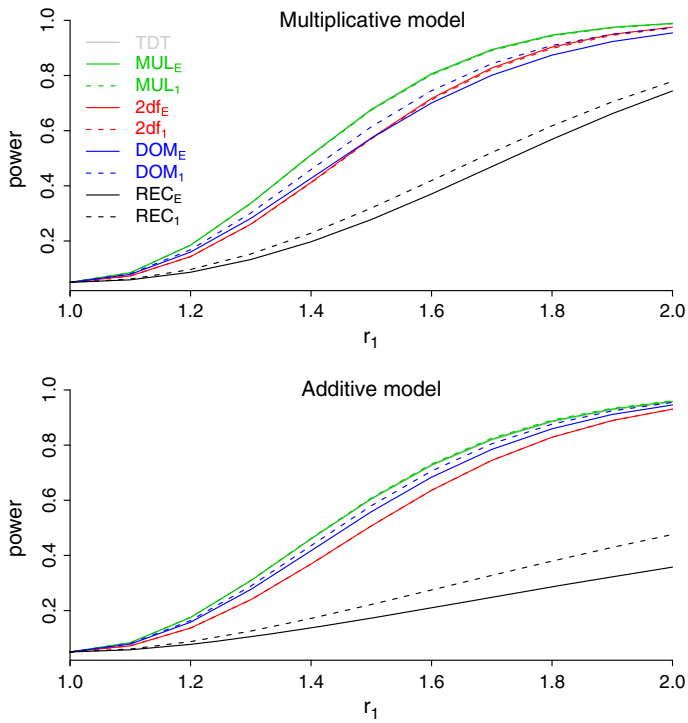

Figure 2. The power of the nine tests: *MUL*, *DOM*, *REC*, and 2*df* tests for each of the two matching strategies, and the TDT test. Because the power curves of the TDT and MUL_E almost overlap each other, it is hard to see the curves of the TDT on the plots. Upper panel: power curves when the true genetic model is multiplicative; lower panel: power curves when the true genetic model is additive.

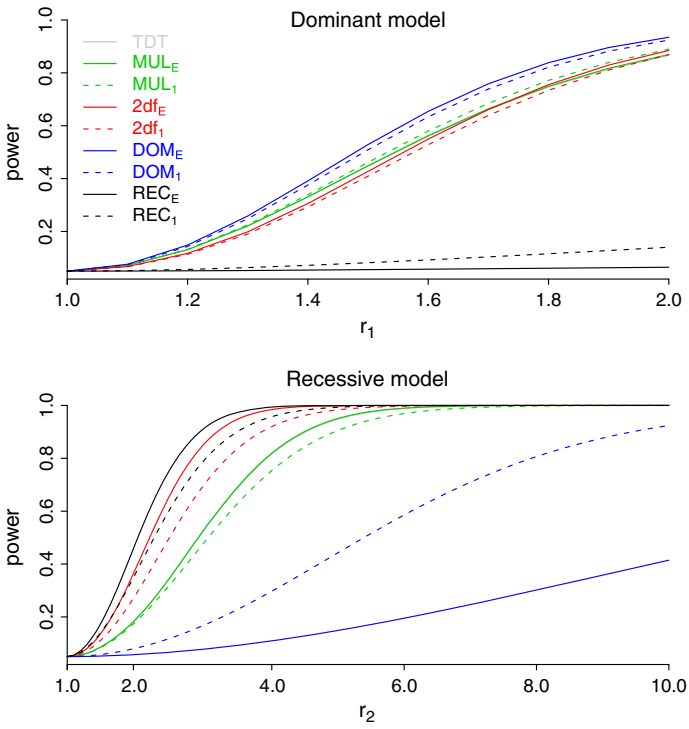

Figure 3. The power of the nine tests: *MUL*, *DOM*, *REC*, and 2*df* tests for each of the two matching strategies, and the TDT test. Because the power curves of the TDT and MUL_E almost overlap each other, it is hard to see the curves of the TDT on the plots. Upper panel: power curves when the true genetic model is dominant; lower panel: power curves when the true genetic model is recessive.

Table III. Range of the power difference (in %) between the 1:3 matching and the 1:1 matching for four testing models (multiplicative, dominant, recessive, 2*df*) and four true genetic models (multiplicative, dominant, recessive, and additive) in the single-locus analysis (A negative value means the 1:1 matching is more powerful, while a positive value means the 1:3 matching is more powerful).

	Testing model			
True model	Multiplicative	Dominant	Recessive	2*df*
Multiplicative	[0, 0.4]	[–4.5, 0]	[–5.1, 0.1]	[0, 0.5]
Dominant	[–2.2, 0]	[0, 2.1]	[–7.6, 0]	[0, 2.2]
Recessive	[0, 6.6]	[–51.9, 0]	[0, 15.0]	[0, 16.3]
Additive	[–0.4, 0]	[–2.2, 0]	[–11.2, 0]	[0, 0]

Table III and the lower panel of Figure 3, the 1:1 matching outperforms the 1:3 matching with a maximum difference of 51.9% (recall that a negative value in Table III implies that the 1:1 matching is more powerful than the 1:3 matching). This demonstrates that the 1:1 matching is more robust against the misspecification of the genetic model.

For the multiplicative test, the 1:1 matching has comparable efficiency to the 1:3 matching for all genetic models, as also shown in Table III. This is especially important because when no prior information regarding the true genetic model is available, one often conducts tests with the multiplicative assumption to prevent substantial power loss from model misspecification.

The results of the HWE test are shown in Figure 4. When the GRRs are multiplicative, the transmission of alleles from parents follows the HWE. As a result, the power of the HWE test is equal to the nominal *p*-value cutoff for both matching strategies. When the GRRs are additive, dominant, or recessive, the 1:3 matching is more powerful than the 1:1 matching. Because those GRRs at a locus can be considered as allelic interactions, the result indicates that the 1:1 matching is less efficient in testing allelic interactions.

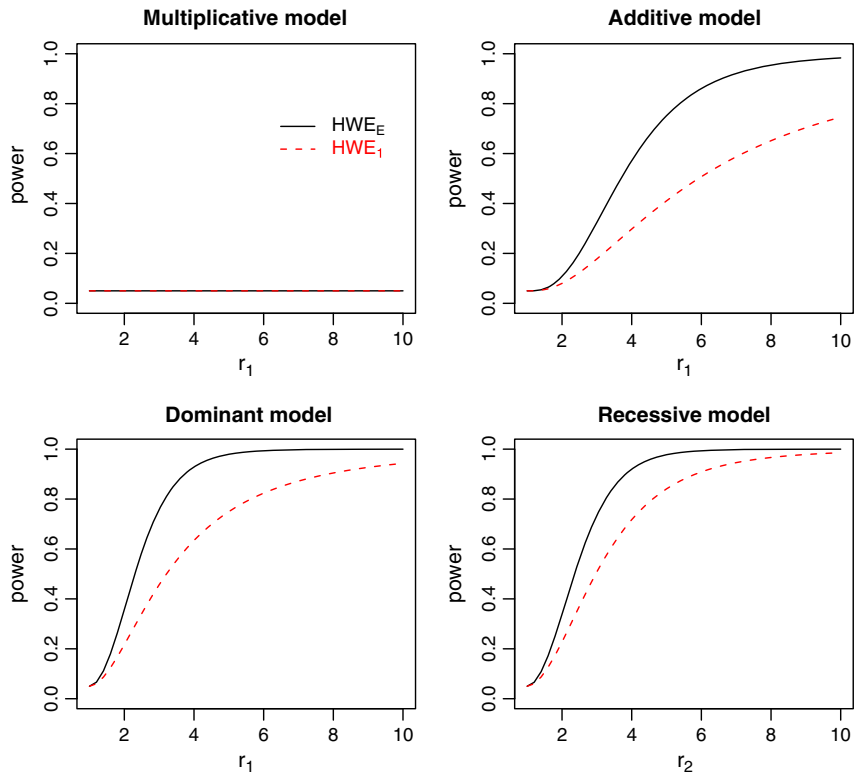

Figure 4. The power of the HWE test for the 1:3 matching (solid line) and the 1:1 matching (dashed line) when the true genetic models are multiplicative, additive, dominant, and recessive, respectively.

3.2. *Results of two-locus models*

The results of the two-locus analysis are shown in Figure 5, and they are similar to those observed in the single-locus analysis. Recall that the main effects in our two-locus analysis are assumed to be multiplicative at both loci. When the testing model agrees with the true main effects, the exhaustive matching is expected to be more powerful than the 1:1 matching; however, we found that under the mul–mul model, the difference in power between the two strategies is quite small, with the maximum difference being less than 0.7% (data not shown). This echoes what we found in the one-locus analysis, that is, for the multiplicative test, the 1:1 matching and the 1:3 matching have comparable efficiency. When the true main effects do not agree with the testing model, the relative efficiency of the 1:1 matching to the exhaustive matching is model-dependent. The 1:1 matching is slightly more powerful in the dom–dom model (the upper left plot of Figure 5) but slightly less powerful in the rec–rec model (the upper right plot of Figure 5). For the gene–gene interaction models, we tested both main effects (shown in the lower left plot of Figure 5) and gene–gene interactions (show in the lower right plot of Figure 5). In detecting main effects, the 1:1 matching is as efficient as the exhaustive matching, with the maximum power reduction being less than 3%. In detecting gene–gene interactions, the 1:1 matching is not as efficient as the exhaustive matching, with the maximum loss being 30.2%.

Intuitively, the exhaustive matching utilizes more information and thus should be more efficient than the 1:1 matching in both single-locus and multilocus analysis. This is largely confirmed in the numerical comparisons above. First, a higher power in detecting main effect was observed with the exhaustive matching method when the model was correctly specified. It seems that most of the time this gain in statistical power was minimal and hardly appreciable in both single-locus and two-locus analyses. This might be partly because not all pseudocontrols constructed under the exhaustive matching provide evidence for the true model. Some pseudocontrols may represent the evidence against the true model,

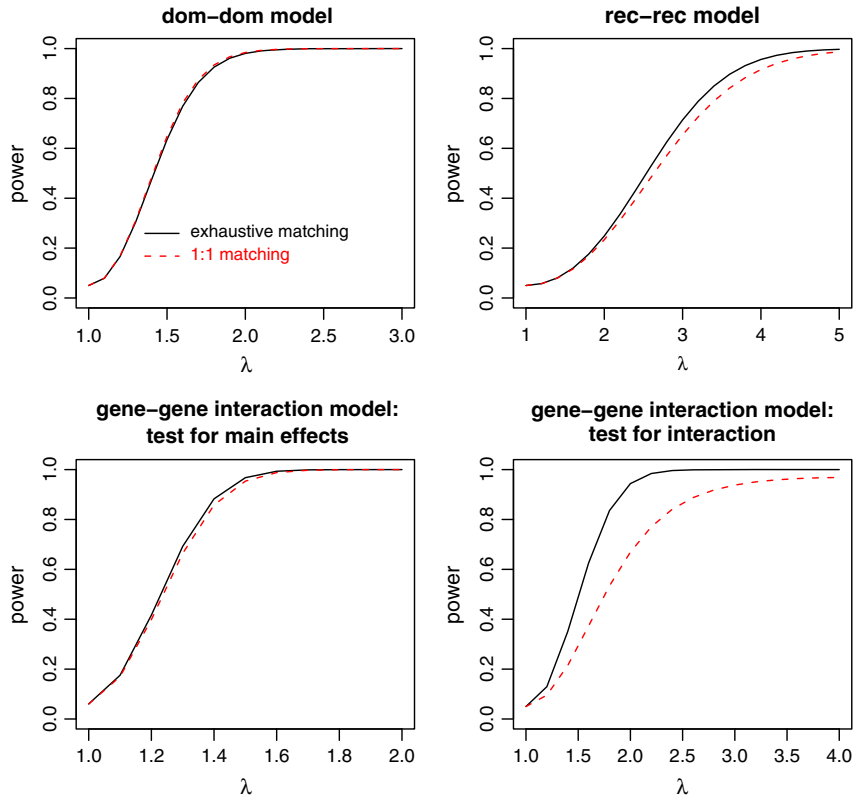

Figure 5. The power computed under two-locus models, for both the 1:1 matching (dashed line) and the exhaustive matching (solid line). The upper two plots show the power of detecting main effects computed under two main-effect-only models: dom–dom and rec–rec, respectively. The power of detecting main effects under the mul–mul model shows no noticeable difference between the 1:1 matching and the exhaustive matching; thus, the results are omitted. The lower plots show the power of detecting main effects and gene-gene interactions, respectively, under the gene-gene interaction model.

offsetting the statistical power in detecting an association. Second, we found that the exhaustive matching is much more powerful than the 1:1 matching in testing for interaction. By the way the 1:1 matching is constructed, it only uses the combination of alleles not passed from the parents as the pseudocontrol; as a result, its power for detecting interaction is thus very much limited.

To summarize, the results of the one-locus and two-locus studies demonstrate that the 1:1 matching generally has comparable power to the exhaustive matching in detecting multiplicative/additive effects in single-locus analysis and main effects in multilocus analysis but not as efficient in detecting interactions, which agrees with our theoretical finding about the equivalence (near equivalence) between the 1:1 matching and the exhaustive matching under additive (multiplicative) GRRs.

3.3. Results of the Crohn's disease dataset

The p-values of the single-locus analysis of the Crohn's disease dataset are shown in Figure 6. Among all tests, the two tests for the multiplicative effect show the highest significance for most SNPs. The difference in p-values between the 1:1 matching and the 1:3 matching is small for the two 2-DOF tests and for the two multiplicative tests. Because the two 2-DOF tests are exactly/approximately identical when the true model is additive/multiplicative, the results here suggest that additivity or multiplicity is a good approximation to the true genetic effects of most SNPs.

Given that the 1:1 matching is efficient in testing main genetic effects in multilocus analysis, we use the 1:1 matching to search SNPs that are jointly associated with Crohn's disease. In the multilocus analysis, we first arbitrarily chose two values for the tuning constant: 1 and 20. The value 1 resulted in selecting 29 out of 103 SNPs and the value 20 resulted in selecting two SNPs. Therefore, it is reasonable to believe that the optimal value of the tuning constant is between 1 and 20. Restricting the tuning constant to be in this range and using the leave-one-out cross-validation, we found that the optimal value of the tuning constant is 3.48. This optimal value led to a conditional logistic regression with nine SNPs, with an overall p-value of 1.6×10^{-7}. Using the final model with the nine selected SNPs, we computed the predicted probability that an offspring or a pseudocontrol is diseased, with the results shown in the upper plot of Figure 7. Using 0.5 as the cutoff to predict whether a child is diseased, the model with the nine selected SNPs correctly predicted 92 out of 127 affected offspring being diseased, which is a 72.4% of correctness. As a comparison, we also computed the percentage of correctly predicted disease status using the most significant SNP. With missing genotypes imputed, the 26 th SNP had the smallest

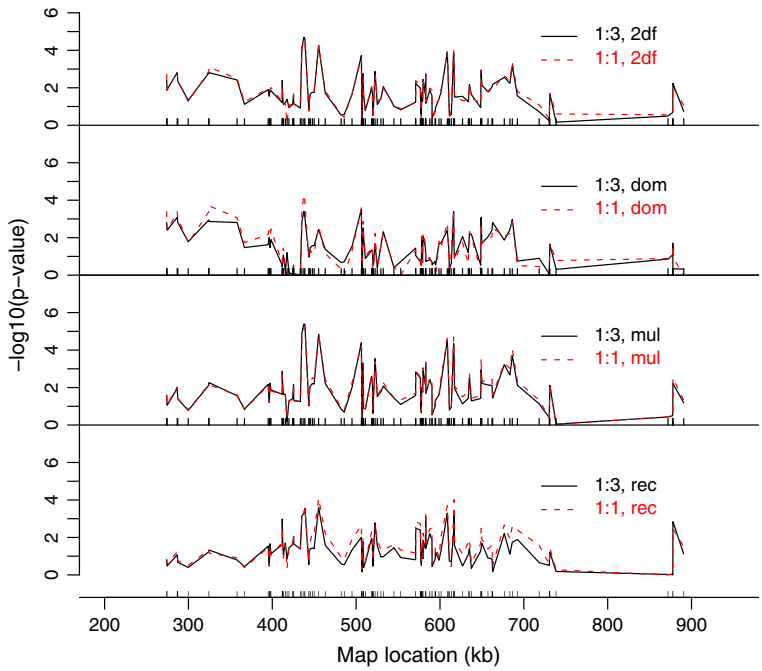

Figure 6. The p-values (on the –log10 scale) of the four single-locus tests for the Crohn's disease data. The four tests are, from top to bottom, the 2*df*, *DOM*, *MUL*, and *REC* tests, respectively. Solid line: the 1:3 matching; dashed line: the 1:1 matching.

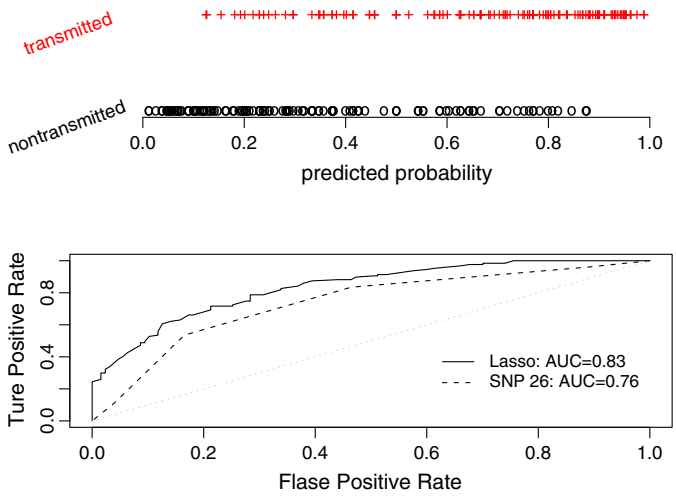

Figure 7. Upper panel: the predicted probabilities for the affected offspring (transmitted) and the pseudosiblings (nontransmitted) being diseased using the penalized multilocus analysis under the 1:1 matching; lower panel: the ROC curves and AUC for SNP 26 selected by the single-locus analysis and for the nine SNPs selected by the penalized multilocus analysis.

p-value (1.61×10^{-6} using 1:1 matching, 1.32×10^{-6} using 1:3 matching) and it provided the best predictability among all individual SNPs, but it only correctly predicted 68 of 127 affected offspring to be diseased, which is 53.5%. The receiver operating characteristic (ROC) curves and the areas under the curves (AUC) are shown in the lower plot of Figure 7. Because we calculated the AUC using the data of the affected offspring and their pseudosiblings, the AUC values we obtained might overestimate what they would be if real case-control data were used. However, the larger AUC of the multilocus analysis nevertheless shows the benefit of using it. These results demonstrate that the 1:1 matching-enabled multilocus analysis provides a improved prediction of disease than the single-locus analysis.

Although the nine selected markers show increased predictability for Crohn's disease than the most significant SNP, we have to interpret the selected markers cautiously. It has been recently shown that untyped rare DNA variations can cause association between a disease and multiple common SNPs [48]. Thus, the selected SNPs might imply that multiple common and rare DNA variations at 5q31 are responsible for Crohn's disease. Note that the region 5q31 has been replicated by several genome-wide association studies, such as [49,50]. Despite the uncertainty in interpreting the detected association based on the Crohn's disease data, our results nevertheless demonstrate that the penalized multilocus method based on the 1:1 matching can improve the power in detecting association. In the future, we plan to evaluate the penalized method with large-scale genome-wide trio studies.

4. Discussion

To summarize, we compared the efficiency of the 1:1 matching to that of the exhaustive matching using both theoretical derivations and numerical computations. Our results show that: (1) the two matching strategies use exactly/approximately the same amount of information under additive/multiplicative GRRs and (2) the 1:1 matching is efficient in detecting main effects in multilocus analysis. We illustrated the usefulness and efficiency of the 1:1 matching in multilocus analysis by applying it to the Crohn's disease data, and we improved the percentage of correctly predicted diseases status by more than 18.9% — from 53.5% achieved by using the most significant SNP to 72.4% achieved by using the nine SNPs selected by the penalized method.

We studied the usefulness of the 1:1 matching in the case-parents design. With minor modifications, the same matching strategy can also be extended to general nuclear families with multiple affected and unaffected offspring. For a family with m affected offspring, we can use the following likelihood function

$$\frac{\prod_{j=1}^{m} \exp(\beta^T g_j)}{\prod_{j=1}^{m} \exp(\beta^T g_j) + \prod_{j=1}^{m} \exp(\beta^T \bar{g}_j)}$$

where g_j is the genotype vector for the j-th affected offspring, $\bar{g}_j$ is the genotype vector constructed by alleles not transmitted to the j-th affected offspring, and β is the vector of the coefficients of the main genetics effects in the conditional logistic regression. This way of creating matched data and constructing the likelihood function corresponds to a permutation test that permutes all siblings within a family, instead of individual siblings in each permutation. Because this permutation maintains the dependence of siblings within the same family [51], the matching we use here also preserves the dependence of siblings and leads to valid tests of association whether linkage presents or not. In addition, we may also include unaffected offspring in our modeling by labeling the alleles of an unaffected offspring as 'non-transmitted' and the alleles not transmitted to the unaffected offspring as 'transmitted' [52]. This has to be used with caution, however, as incorporating the unaffected offspring may lead to either negligible gain in power or even reduced power when the penetrance is not high [53].

Here, we showed that the 1:1 matching has comparable efficiency to the exhaustive matching in many situations. However, we want to point out that our results do not indicate that one should always replace the exhaustive matching with the 1:1 matching. In fact, when testing the association between one locus and a disease, the exhaustive matching (the 1:3 matching) always uses the same or more information from data than does the 1:1 matching. For two unlinked loci, the exhaustive matching is also recommended, because the exhaustive matching (1:15) does not lead to substantially more computational cost than the 1:1 matching. It is when multiple loci are jointly analyzed, especially when the number of jointly analyzed loci is greater than two or when the loci are in linkage, the 1:1 matching provides an efficient, convenient, and easy-to-implement strategy to test main genetic effects. If gene–gene interaction effects are the focus of an analysis, the 1:1 matching is still valid, such as in [26]; however, it might be suboptimal in terms of power. It is clearly worthy of future research efforts to develop statistical methods that can efficiently capture interactions of multiple linked loci using family data.

Appendix

We can arbitrarily label one allele as the risk allele and the other one as the normal allele at each SNP, and define trio types at each SNP in the same way as shown in Table I. Let $n_{t_1,t_2,\ldots,t_L}$ denote the number of trios with type t_l at SNP l for $l = 1, 2, \ldots, L$, where

$$t_l \in \{12, 22, 21, 31, 42, 41, 40, 51, 50, 60\}.$$

The contributions of $n_{t_1,t_2,\ldots,t_L}$ to the exhaustive likelihood and the 1:1 matching likelihood are

$$\left(\frac{r_{g(t_1,t_2,\ldots,t_L)}}{\sum_{g'\in H_E(t_1,t_2,\ldots,t_L)} r_{g'}}\right)^{n_{t_1,t_2,\ldots,t_L}} \quad \text{and} \quad \left(\frac{r_{g(t_1,t_2,\ldots,t_L)}}{\sum_{g'\in H_1(t_1,t_2,\ldots,t_L)} r_{g'}}\right)^{n_{t_1,t_2,\ldots,t_L}},$$

respectively, where $g(t_1, t_2, \ldots, t_L)$ is the genotype of the affected child in a case-parents trio with type $(t_1, t_2, \ldots, t_L)$, $H_1(t_1, t_2, \ldots, t_L)$ denotes the genotypes of the affected child and his or her pseudosibling under the 1:1 matching, and $H_E(t_1, t_2, \ldots, t_L)$ denotes the 4^L genotypes of all possible siblings (including the affected child and the 4^L–1 pseudosiblings) under the exhaustive matching. When the L loci are unlinked, the pseudosibship method implies that the 4^L genotypes in $H_E(t_1, t_2, \ldots, t_L)$ are equally likely.

Under the assumption of additive GRRs, the GRR of a genotype depends only on the number of the risk alleles at each SNP. Thus, for the 1:1 matching, the sum of the GRRs of the affected child and his or her pseudosibling is equal to the sum of the GRRs of the father and the mother. If we let n_l denote the number of risk alleles at SNP l in a pair of parents under trio type $(t_1, t_2, \ldots, t_L)$, we have

$$\sum_{g'\in H_1(t_1,t_2,\ldots,t_L)} r_{g'} = 2 + \sum_{l=1}^{L} n_l\gamma_l.$$

In the exhaustive matching, $H_E(t_1, t_2, \ldots, t_L)$ comprises of all the 4^L combinations of parental alleles. Thus, the sum of the risk alleles of the 4^L genotypes at SNP l is equal to $(4^L/2)n_l$. Thus,

$$\sum_{g'\in H_E(t_1,t_2,\ldots,t_L)} r_{g'} = 4^L + \sum_{l=1}^{L} \frac{4^L}{2} n_l\gamma_l = \frac{4^L}{2} \sum_{g'\in H_1(t_1,t_2,\ldots,t_L)} r_{g'}$$

As a result, we have

$$\left(\frac{r_{g(t_1,t_2,\ldots,t_L)}}{\sum_{g'\in H_E(t_1,t_2,\ldots,t_L)} r_{g'}}\right)^{n_{t_1,t_2,\ldots,t_L}} = \left(\frac{2}{4^L}\right)^{n_{t_1,t_2,\ldots,t_L}} \left(\frac{r_{g(t_1,t_2,\ldots,t_L)}}{\sum_{g'\in H_1(t_1,t_2,\ldots,t_L)} r_{g'}}\right)^{n_{t_1,t_2,\ldots,t_L}}$$

which proves that the likelihood functions for the 1:1 matching and the exhaustive matching are the same.

Acknowledgements

We thank the anonymous reviewers for their careful reading of the manuscript and constructive comments. The author ZY was supported in part by grant NIH/R01 HG004960.

References

1. Li CC. Population subdivision with respect to multiple alleles. *Annals of Human Genetics* 1969; **33**:23–29.
2. Anonymous. Freely associating. *Nature Genetics* 1999; **22**:1–2.
3. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361**:598–604.
4. Thomas DC, Witte JS. Point: Population stratification: A problem for case-control studies of candidate-gene associations? *Cancer Epidemiology Biomarkers & Prevention* 2002; **11**:505–512.
5. Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a propercontrol sample for risk calculations. *Annals of Human Genetics* 1987; **51**:227–233.
6. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics* 1993; **53**:1114–1126.
7. Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 1991; **47**:53–61.
8. Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* 1996; **13**:423–449.
9. Schaid DJ. Likelihoods and TDT for the case-parents design. *Genetic Epidemiology* 1999; **16**:250–260.
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 1993; **52**:506–516.
11. Terwilliger JD, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity* 1992; **42**:337–346.
12. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**:153–157.
13. Wittkowski KM, Liu X. A statistically valid alternative to the TDT. *Human Heredity* 2002; **54**:157–164. DOI: 10.1159/000068840.
14. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *American Journal of Human Genetics* 2002; **70**:124–141. DOI: 10.1086/338007.
15. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *American Journal of Epidemiology* 2002; **155**:478–484. DOI: 10.1093/aje/155.5.478.
16. Thomas DC. *Statistical Methods in Genetic Epidemiology*. New York: Oxford University Press, 2004.
17. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. Transmission/disequilibrium tests using multiple tightly linked markers. *American Journal of Human Genetics* 2000; **67**:936–946. DOI: 10.1086/303073.
18. Allen AS, Satten GA. Statistical models for haplotype sharing in case-parent trio data. *Human Heredity* 2007; **64**:35–44. DOI: 10.1159/000101421.
19. Zhang S, Sha Q, Chen HS, Dong J, Jiang R. Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *American Journal of Human Genetics* 2003; **73**:566–579. DOI: 10.1086/378205.
20. Van der Meulen MA, te Meerman GJ. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genetic Epidemiology* 1997; **14**:915–920. DOI: 10.1002/(SICI)1098-2272(1997)14:6<915::AID-GEPI59>3.0.CO;2-P.
21. Dudbridge F, Koeleman BP, Todd JA, Clayton DG. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *American Journal of Human Genetics* 2000; **66**:2009–2012. DOI: 10.1086/302915.
22. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity* 2003; **56**:18–31. DOI: 10.1159/000073729.
23. Knapp M, Becker T. Family-based association analysis with tightly linked markers. *Human Heredity* 2003; **56**:2–9. DOI: 10.1159/000073727.

24. Fan R, Knapp M, Wjst M, Zhao C, Xiong M. High resolution T2 association tests of complex diseases based on family data. *Annals of Human Genetics* 2005; **69**:187–208. DOI: 10.1046/j.1529-8817.2004.00151.x.
25. Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity* 1998; **48**:67–81.
26. Kotti S, Bickeboller H, Clerget-Darpoux F. Strategy for detecting susceptibility genes with weak or no marginal effect. *Human Heredity* 2007; **63**:85–92. DOI: 10.1159/000099180.
27. Zhang Z, Zhang S, Sha Q. A multi-marker test based on family data in genome-wide association study. *BMC Genetics* 2007; **8**:65. DOI: 10.1186/1471-2156-8-65.
28. McIntyre LM, Martin ER, Simonsen KL, Kaplan NL. Circumventing multiple testing: A multilocus Monte Carlo approach to testing for association. *Genetic Epidemiology* 2000; **19**:18–29.
29. Shi M, Umbach DM, Weinberg CR. Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *American Journal of Human Genetics* 2007; **81**:53–66. DOI: 10.1086/518670.
30. Lee WC. Testing for candidate gene linkage disequilibrium using a dense array of single nucleotide polymorphisms in case-parents studies. *Epidemiology* 2002; **13**:545–551. DOI: 10.1097/01.Ede.0000023392.38744.60.
31. Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics* 1998; **63**:1531–1540.
32. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff RK. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics* 1996; **13**:399–408.
33. Jiang R, Dong J, Wang D, Sun FZ. Fine-scale mapping using Hardy-Weinberg disequilibrium. *Annals of Human Genetics* 2001; **65**:207–219. DOI: 10.1046/j.1469-1809.2001.6520207.x.
34. Song KJ, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine* 2006; **25**:105–126. DOI: 10.1002/Sim.2350.
35. Zheng G, Ng HKT. Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* 2008; **9**:391–399. DOI: 10.1093/biostatistics/kxm039.
36. Yuan M, Tian X, Zheng G, Yang YN. Adaptive Transmission Disequilibrium Test for Family Trio Design. *Statistical Applications in Genetics and Molecular Biology* 2009; **8**. DOI: 10.2202/1544-6115.1451.
37. Loftus EV, Schoenfeld P, Sandborn WJ. The epidemiology and natural history of Crohn's disease in population-based patient cohorts from North America: a systematic review. *Alimentary Pharmacology & Therapeutics* 2002; **16**:51–60. DOI: 10.1046/j.1365-2036.2002.01140.x.
38. Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, Bitton A, Williams CN, Greenberg GR, Cohen Z, Lander ES, Hudson TJ, Siminovitch KA. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *American Journal of Human Genetics* 2000; **66**:1863–1870. DOI: 10.1086/302913.
39. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, Kulbokas EJ, O'Leary S, Winchester E, Dewar K, Green T, Stone V, Chow C, Cohen A, Langelier D, Lapointe G, Gaudet D, Faith J, Branco N, Bull SB, McLeod RS, Griffiths AM, Bitton A, Greenberg GR, Lander ES, Siminovitch KA, Hudson TJ. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics* 2001; **29**:223–228. DOI: 10.1038/ng1001-223.
40. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nature Genetics* 2001; **29**:229–232. DOI: 10.1038/ng1001-229.
41. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* 1996; **58**:267–288.
42. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009; **25**:714–721. DOI: 10.1093/bioinformatics/btp041.
43. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology* 2010; **34**:879–891. DOI: 10.1002/gepi.20543.
44. Croiseau P, Cordell HJ. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC Proceedings* 2009; **3**(Suppl 7). DOI: 10.1186/1753-6561-3-s7-s61.
45. Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 2010; **26**:2375–2382. DOI: 10.1093/bioinformatics/btq448.
46. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 2010; **52**:70–84. DOI: 10.1002/bimj.200900028.
47. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* 2009; **84**:210–223. DOI: 10.1016/j.ajhg.2009.01.005.
48. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *Plos Biology* 2010; **8**. DOI: 10.1371/journal.pbio.1000294.
49. Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, Bradley WE, Croteau P, Nguyen-Huu Q, Segal J, Debrus S, Allard R, Rosenstiel P, Franke A, Jacobs G, Nikolaus S, Vidal JM, Szego P, Laplante N, Clark HF, Paulussen RJ, Hooper JW, Keith TP, Belouchi A, Schreiber S. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proceedings of the National Academy of Sciences of the United States of America* 2007; **104**:14747–14752. DOI: 10.1073/pnas.0706645104.
50. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S,

Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ, Consortium NIG, Consortium B-FI, Control WTC. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* 2008; **40**:955–962. DOI: 10.1038/Ng.175.

51. Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. *American Journal of Human Genetics* 1997; **61**:439–448. DOI: 10.1086/514860.
52. Guo CY, Lunetta KL, DeStefano AL, Ordovas JM, Cupples LA. Informative-transmission disequilibrium test (i-TDT): combined linkage and association mapping that includes unaffected offspring as well as affected offspring. *Genetic Epidemiology* 2007; **31**:115–133. DOI: 10.1002/gepi.20195.
53. Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics* 1999; **65**:1170–1177. DOI: 10.1086/302577.