Genotype-Based Matching to Correct for Population Stratification in Large-Scale Case-Control Genetic Association Studies

Weihua Guan**, Liming Liang**, Michael Boehnke, and Gonçalo R. Abecasis*

Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, Michigan

Genome-wide association studies are helping to dissect the etiology of complex diseases. Although case-control association tests are generally more powerful than family-based association tests, population stratification can lead to spurious disease-marker association or mask a true association. Several methods have been proposed to match cases and controls prior to genotyping, using family information or epidemiological data, or using genotype data for a modest number of genetic markers. Here, we describe a genetic similarity score matching (GSM) method for efficient matched analysis of cases and controls in a genome-wide or large-scale candidate gene association study. GSM comprises three steps: (1) calculating similarity scores for pairs of individuals using the genotype data; (2) matching sets of cases and controls based on the similarity scores so that matched cases and controls have similar genetic background; and (3) using conditional logistic regression to perform association tests. Through computer simulation we show that GSM correctly controls false-positive rates and improves power to detect true disease predisposing variants. We compare GSM to genomic control using computer simulations, and find improved power using GSM. We suggest that initial matching of cases and controls prior to genotyping combined with careful re-matching after genotyping is a method of choice for genome-wide association studies. *Genet. Epidemiol.* 33:508–517, 2009. © 2009 Wiley-Liss, Inc.

Key words: population stratification; genome-wide association; genetic similarity

Contract grant sponsor: National Institutes of Health; Contract grant numbers: HG000376; HG02651; HL084729.

*Correspondence to: Gonçalo R. Abecasis, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor MI 48109-2029. E-mail: goncalo@umich.edu

**These authors contributed equally to this work.

Received 27 November 2008; Revised 27 November 2008; Accepted 10 December 2008

Published online 23 January 2009 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.20403

INTRODUCTION

With the success of the International HapMap Project [The International HapMap Consortium, 2007], a dense set of single nucleotide polymorphisms (SNPs) throughout the human genome is now available for genetic studies of complex diseases, and many genome-wide association studies are being undertaken and published [Klein et al., 2005; Maraganore et al., 2005; Cheung et al., 2005; Sladek et al., 2007; Scott et al., 2007; Saxena et al., 2007; Zeggini et al., 2007].

Although case-control association tests are in principle more powerful for detecting disease variants than family-based association tests, population stratification can lead to spurious disease-marker association or mask true association [Li, 1972]. In genome-wide association studies, thousands of samples are typically used to ensure adequate power to identify disease predisposing variants, making it difficult to guarantee genetic homogeneity of the sample [Freedman et al., 2004]. Ancestry information on the sampled individuals may be unavailable to the researchers, and even when available, may not fully specify the underlying population genetic structure, due to vague definitions of ancestry groups and imperfect accuracy of self-report information.

Several methods have been proposed to adjust for the possible confounding effects of population substructure. Family-based association tests, such as the transmission/disequilibrium test [Spielman et al., 1993], assess the transmission of alleles from parents to affected offspring. Comparisons are made within parent-offspring trios, and the resulting association test is immune to potential genetic heterogeneity between families. However, collecting trios can be difficult and expensive, and may simply be impractical for late-onset diseases. For unrelated case-control samples, approaches have been proposed to adjust the standard χ^2 contingency test statistics according to a non-central χ^2 distribution [Devlin and Roeder, 1999; Gorroochurn et al., 2006], to infer population structure [Pritchard et al., 2000], or to cluster the similarity estimates into several components [Zhang et al., 2002]. A few more recent approaches [Price et al., 2006; Epstein et al., 2007; Kimmel et al., 2007; Luca et al., 2008] focus specifically on genome-wide association studies.

In this article, we propose a different approach, genetic similarity score matching (GSM), to correct population stratification using individual-based matching rather than clustering. The huge amounts of data in genome-wide association studies have the potential to provide extremely accurate matching of individuals who share similar ancestries. We match cases with controls based on genetic (dis)similarity scores calculated from the genotype data available in a genome-wide association study or a large-scale candidate gene study and test the resulting matched

© 2009 Wiley-Liss, Inc.

sets for disease-marker association by conditional logistic regression. This matching-association framework builds on our previous work [Guan et al., 2005] and is similar to that of Luca et al. [2008]. Luca et al. [2008] derive the dissimilarity (distance) scores based on principal components of the variance matrix of genotypes, while our approach obtains the dissimilarity scores based on identity-by-state (IBS) measures. Simulations show that GSM results in false-positive rates at the desired nominal level while retaining high power to detect disease-associated markers. We find that with large-scale association data, the calculated genetic similarity scores differentiate subpopulations well, and that matching can be done with high accuracy even for samples that are mixtures of genetically similar populations. We further demonstrate that when population stratification is present, association tests based on GSM-matched case-control data can have a higher power than those that rely on either the standard trend test or the genomic-control method.

METHODS

OUTLINE

GSM includes three basic components:

(1) *Genetic similarity score*: We calculate genetic similarity scores between pairs of cases and controls across all loci. Large scores should reflect pairs with similar genetic backgrounds.
(2) *Matching*: Based on the matrix of similarity scores calculated in (1), we conduct optimal full matching [Rosenbaum, 2002] which groups one case with one or more controls, or one control with one or more cases to maximize the overall similarity of matched cases and controls.
(3) *Association tests*: We use conditional logistic regression to assess the association between candidate markers and disease status. For ease of exposition, we consider here only single marker association tests, but other genetic or environmental factors can be easily incorporated into the regression.

GENETIC SIMILARITY SCORE

We define a genetic similarity score for a pair of individuals which measures the degree of similarity of their genotype data. Individuals with similar genetic backgrounds will generally have higher scores. For simplicity, we consider M biallelic genetic markers each with alleles "A" and "a"; the scores can easily be generalized to multiallelic markers. We consider three similarity scores.

The first score calculates the proportion of marker alleles shared identical by state (IBS). If IBS_k is the number of alleles shared at marker k (Table I), then

$$S_{\mathrm{IBS}} = \frac{1}{2M^*}\sum_{k=1}^{M^*} \mathrm{IBS}_k, \tag{1}$$

where $1 \le M^* \le M$ is the number of markers that are successfully genotyped in both individuals.

While S_{IBS} has the virtue of simplicity, we may want to allow different markers to make different contributions to measure similarity. For example, we may wish to weight sharing a rare allele more strongly than sharing a common allele. We define our second score as

TABLE I. Values of IBS_k and $\mathrm{IBS}_{k,i}$ for calculation of similarity scores

Genotype pair	IBS_k	$\mathrm{IBS}_{k,A}$	$\mathrm{IBS}_{k,a}$
aa aa	2	0	2
aa Aa	1	0	1
aa AA	0	0	0
Aa Aa	2	1	1
Aa AA	1	1	0
AA AA	2	2	0

$$S_{\mathrm{freq}} = -\frac{1}{2M^*}\sum_{k=1}^{M^*}\sum_{i\in\{A,a\}} \mathrm{IBS}_{k,i}\cdot\log(q_{k,i}), \tag{2}$$

where $q_{k,i}$ is the frequency of allele i at marker k, and $\mathrm{IBS}_{k,i}$ is the number of copies of allele i at marker k shared by the pair of individuals (Table I). We can estimate $q_{k,i}$ using our sample or from the results of previous studies.

In a random mating population, markers are expected to follow Hardy-Weinberg Equilibrium (HWE). When population subdivision is present, tests of HWE tend to be significant owing to excess homozygosity. Our third score takes advantage of this by weighting markers based on their one-sided (excess homozygosity) HWE test P-value p_k [Wigginton et al., 2005]:

$$S_{\mathrm{HWE}} = -\frac{1}{2M^*}\sum_{k=1}^{M^*} \mathrm{IBS}_k\cdot\log(p_k). \tag{3}$$

To avoid the impact of genotyping error that may lead to strong deviation from HWE, we exclude the markers that fail quality control; practically speaking, this might mean using markers with HWE P-value satisfying $P > 10^{-6}$.

As an example, suppose three cases and three controls are genotyped at three loci, as listed in Table II. Then the similarity scores S_{IBS} are as listed in Table III.

For matching, we may use all genotyped markers, or a selected subset. For example, we might pick the markers with the smallest P-values in an HWE test for excess homozygosity, excluding those that fail quality control, in the hope that the selected markers provide maximal information about population stratification in the sample. Further, to avoid selecting markers which are highly correlated, we might choose at most one marker in every n-marker window or per linkage disequilibrium group.

In our analyses, matching relies on a transformed dissimilarity score, defined as

$$D_{ij} = f(S_{ij}) = \left(\frac{\max - S_{ij}}{\max - \min}\right)^2, \tag{4}$$

where $\max = \max_{i,j} S_{ij}$ and $\min = \min_{i,j} S_{ij}$, the maximum and minimum similarity scores among all case-control pairs.

MATCHING

We use the chosen (dis)similarity score to identify optimal matches between cases and controls. The simplest matching scheme is a 1:1 match in which each case is matched to a unique control. This approach is widely used

TABLE II. Example genotypes

Cases		Controls	
Individual	Genotype	Individual	Genotype
1	aa, aa, AA	4	aa, aa, Aa
2	aa, aa, Aa	5	Aa, AA, aa
3	AA, AA, aa	6	AA, AA, aa

TABLE III. Similarity (dissimilarity) scores for individuals in Table II

	Controls		
Cases	4	5	6
1	5/6 (1/36)	1/6 (25/36)	0 (1)
2	1 (0)	2/6 (16/36)	1/6 (25/36)
3	1/6 (25/36)	5/6 (1/36)	1 (0)

but has obvious drawbacks. For example, when the numbers of cases and controls are not equal, some subjects must be discarded, resulting in a loss of information. Further, samples from various subpopulations often are not equally represented among the cases and controls, leading to forced mismatches if only 1:1 matching is allowed

Instead, we consider an optimal matching approach that minimizes the total dissimilarity score:

$$T = \sum_{s=1}^{S} \sum_{i \in A_s, j \in B_s} D_{ij}.$$

Here, A_s and B_s are the sets of cases and controls in a matched set s, and S is the total number of matched sets. It has been shown that an optimal solution to this minimization problem is a full matching, in which each matched set contains one case and one or more controls, or one control and one or more cases, that is, a 1:m or m:1 matching [Rosenbaum, 1991]. Given n cases and n controls, the summation can in principle contain as few as n terms for 1:1 matching to as many as $2(n-1)$ terms for 1:$n-1$ and $n-1$:1 matching. Since large sets result in larger numbers of terms, optimization tends to favor small matched sets. This helps mitigate any potential power loss due to unbalanced matching, i.e., 1:m or m:1 matching with $m \gg 1$ (see section "Discussion").

The problem of minimizing the total dissimilarity score T is analogous to the classic minimum cost flow (MCF) problem in computer science [Rosenbaum, 1991; Hansen, 2004; Hansen and Klopfer, 2006] (Appendix A), and can be solved using the RELAX-IV algorithm [Bertsekas and Tseng, 1994; Frangioni and Manca, 2006]. Given precalculated dissimilarity scores and an upper bound on m, determining the optimal matched set takes on the order of $n^3 \log n$ operations, where n is the total number of subjects. The choice of parameter m constrains the size of matched sets and is somewhat arbitrary; we typically require $m \leq 5$ when numbers of cases and controls are comparable (see section "Discussion"). Prior to matching, we may exclude a few individuals with maximum similarity scores that are extremely small (this is the *caliper* parameter recommended by Hansen and Klopfer, 2006). In datasets including ~2,000 individuals, the matching typically takes <1 min on a modern PC workstation.

To continue with the previous example, we calculate the dissimilarity scores in Table III, and perform both 1:1 matching and optimal matching. In 1:1 matching, the best match yields three pairs: (1, 4), (2, 5), and (3, 6). The total dissimilarity score is $1/36+16/36+0 = 17/36$. In contrast, the optimal full match has two matched sets: (1, 2, 4) and (3, 5, 6). The matched sets include 4 case-control pairs: (1, 4), (2, 4), (3, 5), and (3, 6). The total dissimilarity score is $1/36+0+1/36+0 = 2/36$. In this example, the individuals within group (1, 2, 4) and (3, 5, 6) are similar to each other, and less similar to the individuals in the other group. Full matching offers an obvious matching advantage over 1:1 matching here. In the general case, full matching is guaranteed to produce a total dissimilarity score that is no greater than that obtained using 1:1 matching.

CONDITIONAL LOGISTIC REGRESSION

Once matching is done, a natural choice for matched-set analysis is to use conditional logistic regression to test for disease-marker association. We employ an additive model for association by assigning values of 0, 1, and 2 to genotypes AA, Aa, and aa, respectively. Other genotyping coding schemes could be considered, corresponding for example to dominant, recessive, or general models. The regression can easily incorporate genotype, covariate, and interaction effects.

In a genome-wide association scan, we apply conditional logistic regression analysis to each marker separately. The multiple testing problem can be addressed using Bonferroni correction, permutation, or false-discovery rates.

SIMULATION

We simulated case-control data influenced by genotypes at a disease locus with alleles D and d, under six additive disease models (Table IV). We assumed sampling from a population that consisted of two subpopulations. We randomly sampled 500 cases and 500 controls from this mixed population. For each model, the relative risk (RR) of the predisposing variant allele is set to be the same in different populations. For models 1 and 2, the disease prevalences $K_1 = K_2$ and predisposing variant allele frequencies $q_1 = q_2$; these models represent the scenario of no population stratification. For models 3 and 4, $K_1 < K_2$, creating population stratification in the simulated data. For models 5 and 6, $K_1 < K_2$ and $q_1 \neq q_2$. For model 5, the first population has lower prevalence but higher predisposing variant allele frequency ($K_1 = 0.07$, $q_1 = 0.55$), than the second population ($K_2 = 0.13$, $q_2 = 0.45$). For model 6, the population with higher prevalence also has higher predisposing variant allele frequency ($K_2 = 0.13$, $q_2 = 0.55$) than the other population ($K_1 = 0.07$, $q_1 = 0.45$). For each model, we simulated 500 datasets.

We simulated autosomal SNPs using GENOME, a coalescent-based simulator [Hudson, 1983, 1990; Donnelly and Tavaré, 1995; Liang et al., 2007]. Assuming discrete generations, GENOME simulates the genealogy of a sample of sequences. As the algorithm proceeds backwards in time, coalescence, recombination, and migration events are simulated. Multiple events can occur in the

TABLE IV. Characteristics of simulated disease models: samples drawn from two subpopulations in 1:1 ratio

	Population 1			Population 2		
Model	K_1	p_1	RR_1	K_2	p_2	RR_2
1	0.10	0.5	1.6	0.10	0.5	1.6
2	0.10	0.2	1.6	0.10	0.2	1.6
3	0.07	0.5	1.6	0.13	0.5	1.6
4	0.07	0.2	1.6	0.13	0.2	1.6
5	0.07	0.55	1.6	0.13	0.45	1.6
6	0.07	0.45	1.6	0.13	0.55	1.6

K_i, disease prevalence in population i; p_i, predisposing variant allele frequency in population i; RR_i, relative risk of the predisposing variant allele in population i.

same generation. We set the effective population size as 10,000, the recombination rate as 10^{-8} per base pair, and the mutation rate as 10^{-9} per base pair, assuming the infinite-site mutation model [Kimura, 1969]. We set the rate of migration between subpopulations to 0.0025 per individual per generation, which resulted in a distribution of allele frequency differences similar to that observed when comparing HapMap Han Chinese (HCB) and Japanese (JPT) samples (www.hapmap.org). In particular, the mean allele frequency difference between the two simulated populations is 0.0470, compared to 0.0477 between the HCB and JPT samples. The simulated genome scans surveyed autosomal genomes of ~2,866 Mb composed of 22 chromosomes, whose lengths approximate the actual lengths of the human autosomes (NCBI build 33, www.ncbi.nlm.nih.gov/genome/seq/). We randomly selected 300,000 SNPs with minor allele frequencies >0.05, and choose a disease liability locus with the desired allele frequencies.

To calculate the similarity scores, we used 10,000 markers with the smallest one-sided HWE *P*-values, choosing no more than one marker from each 10-marker window. We set the maximum size of matched groups (*m*) to 6. We compared the type I error and power of GSM, the trend test, genomic control, and EIGENSTRAT for each simulated setting. Given that the simulated samples were drawn from two subpopulations, we used the first principal component to adjust for stratification in EIGENSTRAT; using additional principal components gave similar results. The estimated type I error rates are the proportion of simulated SNPs in which the association test *P*-value is less than the nominal value 10^{-6}, a significance threshold similar to that typically used in genome-wide scans. In this evaluation of type I error rates, we only considered SNPs that were effectively unlinked to the disease locus. We calculated power as the proportion of simulated replicates where the empirical *P*-value is $<10^{-6}$ at the disease locus using a threshold obtained by inspection of test statistics at the null loci.

BIPOLAR DATA

We applied GSM to genome-wide association data from the Pritzker Consortium bipolar study (unpublished data). We selected 717 independent bipolar I European American cases and 779 independent European American controls from NIMH Human Genetics Initiative (www.nimhgenetics.org); controls were carefully matched to cases by self-reported ethnicity prior to genotyping. In addition, we downloaded genotype data on 3,182 independent European American controls from Illumina iControlIDB database (www.illumina.com/pages.ilmn?ID = 231). All individuals were genotyped using the Illumina HumanHap550 BeadChip; 505,796 autosomal SNPs passed quality-control criteria in the Prtizker bipolar study: (1) HWE *P*-values$>10^{-5}$; (2) genotype call rate $>95\%$; and (3) no more than 1 non-Mendelian inheritance or inconsistency among 15 father-mother-offspring trios and 30 duplicate samples. Of these, we excluded 1,632 SNPs due to allele frequency differences >0.05 between the Illumina and Pritzker control samples. We applied GSM and trend tests for association on the Pritzker samples alone and then on the combined Pritzker and Illumina samples. In GSM, we used the 100,000 markers that passed quality control and have the smallest *P*-values from the one-sided HWE test to calculate the similarity scores. Given the relatively large control:case ratio of $3{,}961/717 \approx 5.5$, we set the upper limit of the group sizes (*m*) to 30.

RESULTS

SIMILARITY SCORE PERFORMANCE IN HAPMAP

We first examined the performance of our similarity scores in the HapMap dataset. We calculated our three similarity scores for all pairs of the 89 independent Han Chinese (CHB) and Japanese (JPT) individuals in the HapMap sample, using 100,000 HapMap phase I autosomal SNPs with MAF>0.05, selected based on one-sided HWE test *P*-values of 4.3×10^{-6} to 0.11. In Figure 1, we showed plots from using multidimensional scaling on the similarity score matrices. All three scores showed good separation between the two populations, except for one JPT individual residing in between the two clusters in the plots. The same individual is at a similar position in principal component analysis (PCA) when plotting the first two principal components. While S_{IBS} and S_{freq} provided similar separation, S_{HWE} provided less separation with that JPT individual much closer to the CHB cluster instead of the JPT cluster. The relatively poorer performance of S_{HWE} arises because of the heavy weighting of the small subset of markers with very small *P*-values from the one-sided HWE test, even after we have excluded markers with HWE *P*-value $<10^{-6}$.

Our experiences in simulations and real data (unpublished results) suggest that S_{freq} may perform slightly better than S_{IBS} in matching the samples. In the following simulations and analyses, we report results using S_{freq} as our measure of genetic similarity. Although the *P*-values from one-sided HWE test may not be the best weights for the similarity score as in S_{HWE}, they can still be employed to select a subset of markers for the score computation. In doing so we assume that markers with small HWE *P*-values but still passing quality control provide more information about population heterogeneity than randomly selected markers. In the following analyses, the matching is usually based on a subset of markers (10,000–100,000 markers) which had the smallest *P*-values from one-sided HWE test among those passing quality control filters.

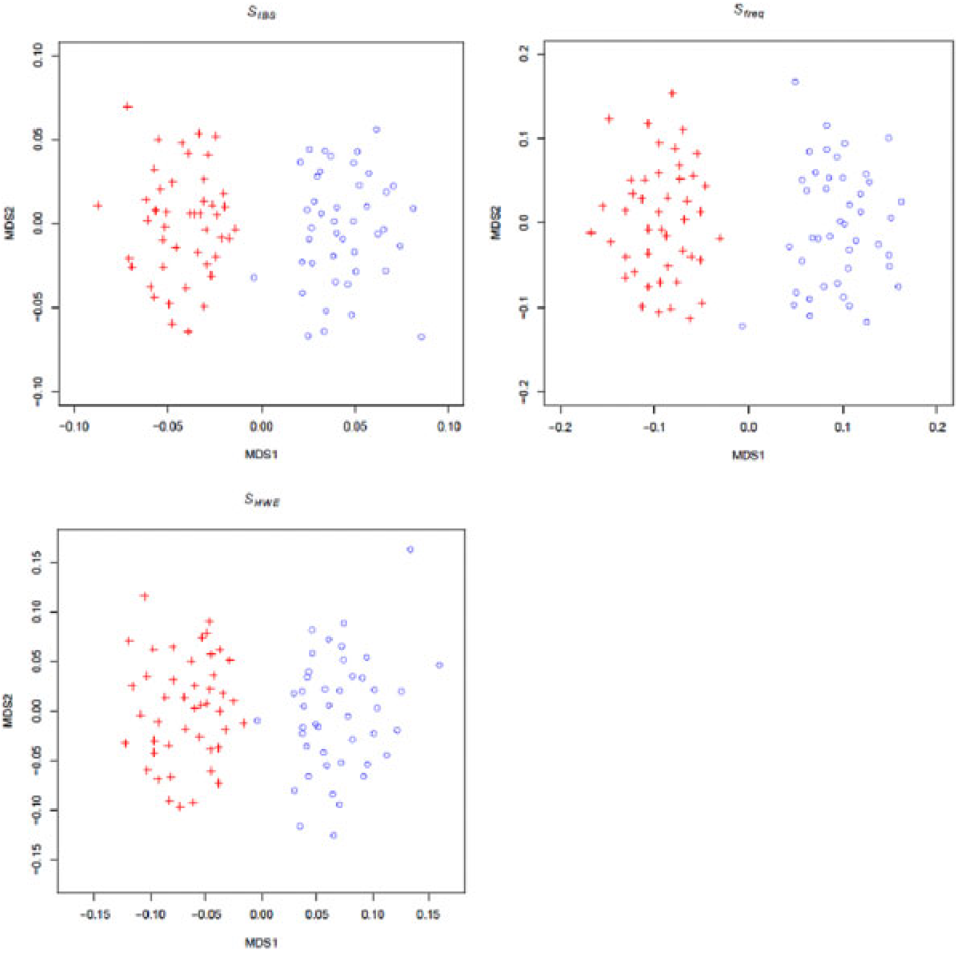

Fig. 1. Multidimensional scaling plots using dissimilarity scores as distance measure (calculated from 100,000 SNPs) for Han Chinese (HCB) and Japanese (JPT) HapMap samples. Red, HCB; blue, JPT.

FALSE-POSITIVE RATE AND POWER

For the six simulation models, mismatch rates are calculated as the proportion of individuals from population 1 matched to individuals from population 2. The minimal degree of mismatch in the simulations (Table V) suggests accurate matching given the similarity measures and numbers of markers used.

In the absence of population stratification (models 1 and 2), all three methods give false-positive rates close to the nominal value of 10^{-6}. The power of our GSM method is typically ~2% lower than the trend test and genomic control, assumedly due to the unnecessary grouping of samples. When population stratification is present (models 3–6), the type I error rate of the trend test is ~30 times greater than the nominal value, while GSM and genomic control maintain the type I errors at or lower than the nominal value. Using empirical type I error rates, the power of the trend test is equal to that of genomic control, but significantly lower than that of GSM for models 3–4. For models 5 and 6, where population stratification is present, the variation of disease variant frequency may mask the association (model 5) or increase the power to detect association (model 6). For model 5, power of the trend test and genomic control drop ~30% compared to model 3, while GSM maintains the same level of power. For model 6, although the type I error is inflated, the trend test has adjusted power comparable to that of GSM. EIGENSTRAT has power similar to GSM in all simulation settings examined.

We also compared the frequency with which the disease variant is the most strongly associated marker, or among the most strongly associated 10, 100, and 1,000 markers, in the trend test or GSM (Fig. 2). The results are consistent with the observations above. In the absence of population stratification (models 1 and 2), the trend test identifies the disease variant slightly more frequently than GSM. When population stratification is present, GSM picks the correct disease variant more frequently for models 3–5. For model 6, GSM picks the correct disease variant almost as frequently as the trend test.

BIPOLAR DATA

We first applied standard trend tests to the Pritzker bipolar case and control samples. The estimated genomic control variance inflation factor λ of the test statistics was

TABLE V. Average false-positive rate and power of GSM, trend test (χ^2), and genomic control (GC) given 500 cases and 500 controls, 300,000 SNPs with MAF >0.05, significance level $=10^{-6}$

			Average false-positive rate ($\times 10^{-6}$)				Power[b]		
Setting	Mismatch (%)	λ[a]	GSM	χ^2	GC	EIGEN	GSM	GC	EIGEN
1	0	1.01	1.08	1.29	1.19	0.93	0.80	0.82	0.82
2	0	1.01	1.10	1.16	1.10	0.97	0.55	0.56	0.56
3	0.016	1.39	1.17	31.8	0.73	1.03	0.75	0.53	0.76
4	0.015	1.38	1.15	30.7	0.47	1.07	0.54	0.28	0.55
5	0.010	1.37	1.14	31.2	0.64	0.90	0.72	0.22	0.72
6	0.010	1.38	1.09	33.0	0.66	0.87	0.79	0.78	0.81

[a] The global correction parameter in genomic control (GC), averaged over simulation replicates.

[b] Power adjusted for the nominal false-positive rates.

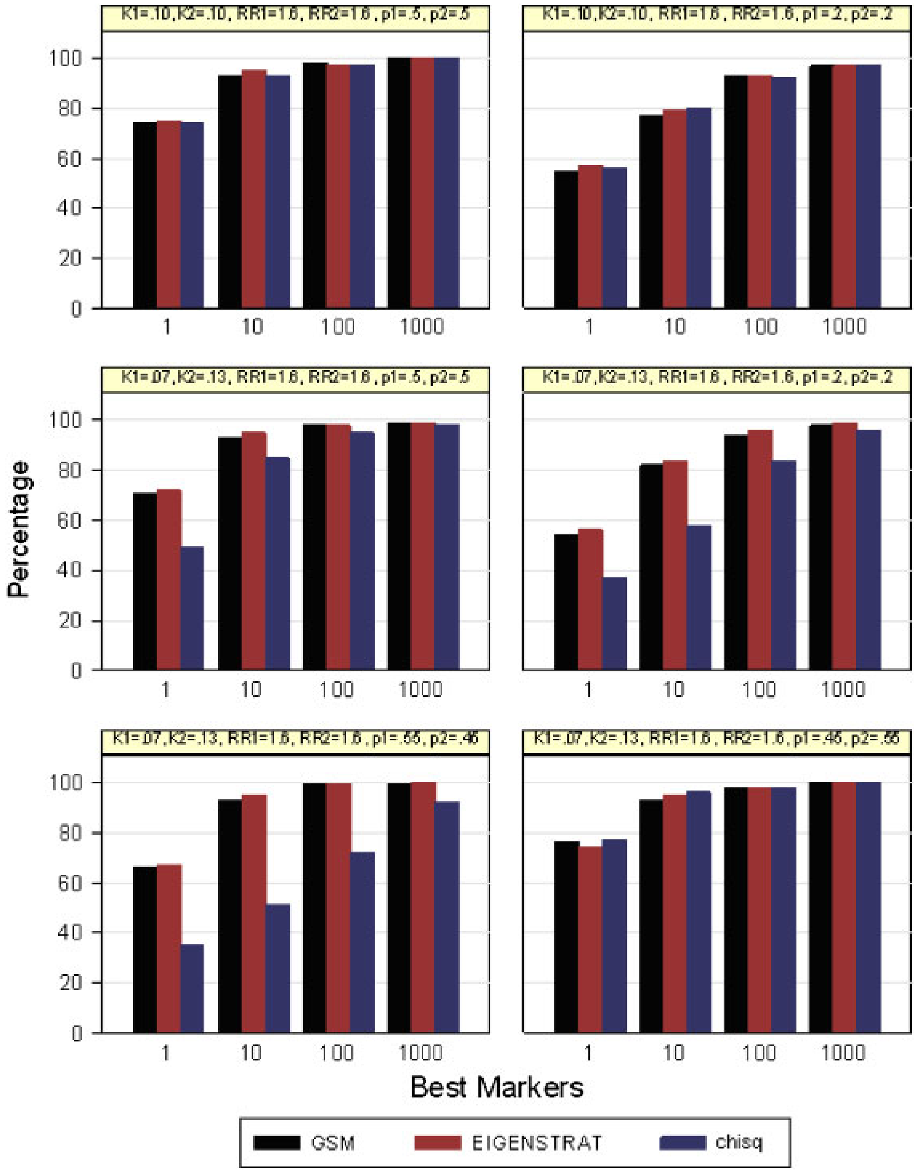

Fig. 2. The frequencies of disease predisposing variant being identified among the best markers by similarity score matching method (GSM), EIGENSTRAT, and trend test (χ^2).

1.03, close to the expected value of 1 when there is no population stratification [Devlin and Roeder, 1999], arguing that the matching based on self-reported ethnicity resulted in a sample with only limited population stratification. Applying GSM reduced the estimated λ slightly to 1.02. However, when we added the Illumina control samples to the analysis, the estimated λ from standard trend tests became 1.51, indicative of strong population stratification between the cases and controls. We then applied our GSM method on the combined samples, excluding one Illumina control sample that had a noticeably high similarity score with one Pritzker case sample ($S_{\mathrm{IBS}} = 0.85$), consistent with a first-degree relationship. Using GSM, the estimated λ dropped to 1.072 when we used S_{freq} as our similarity measure and 1.088 using S_{IBS}, suggesting that GSM using either score provided good correction for the stratification problem. Using $\mathrm{S}_{\mathrm{freq}}$, each of the 712 cases was matched to one or more controls (i.e., 1:*m* matching only): 316 cases were matched to 1 control, 207 cases to 2–5 controls, 79 cases to 6–10 controls, and 115 cases to 10–30 controls. To check the appropriateness of setting the maximum number of controls (*m*) at 30, we repeated our analysis by changing *m* to 10 or 50, resulting in estimated λ values of 1.23 and 1.067, respectively. This suggests that some controls may be matched to dissimilar cases when we only allow up to 10 controls per case, while increasing *m* from 30 to 50 resulted in little improvement on the matching. Since the combined sample contains many more controls than cases, we considered removing some controls with relatively high dissimilarity by restricting the total number of controls to be matched from 3,960 to 3,500, and the estimated λ dropped slightly to 1.065. We also repeated the matching using 50,000 markers instead of 100,000, and in this setting the estimated λ increased slightly to 1.086, as expected

As a comparison, we also applied EIGENSTRAT and another principal component-based method (Luca et al. [2008], GEM) to the bipolar data, using 10 principal components. Without removing any potential outliers, EIGENSTRAT gave an estimated λ of 1.074, comparable to our results. GEM removed 132 samples as outliers and gave a slightly better estimated λ of 1.063. When we applied our method to the same set of samples used in GEM, we obtained an estimate λ of 1.065. Although the removal of these samples decreased the inflation of type I error rates, its impact on power requires further investigation.

DISCUSSION

Population stratification, which can result in high false-positive rates and mask true associations, poses a potential problem for case-control association studies. In this article, we propose GSM, a practical approach to correct for population stratification for large-scale association studies that uses information at thousands of genotyped genetic markers to group case and control subjects according to their similarity. Simulation studies show that GSM can control the false-positive rates in the presence of population substructure, while maintaining power to detect disease loci.

GSM is computationally efficient. The computational time for similarity score calculation is linear in the number of markers used and in the number of all case-control pairs, and the time for matching is approximately cubic in the number of individuals.

We have compared the performance of GSM to the commonly used genomic control method [Devlin and Roeder, 1999]. Genomic control assumes that a scaled test statistic (dividing the standard test statistic by a global correction factor λ) has an approximate central χ^2 distribution. When stratification is modest, the genomic control procedure is able to control the false-positive rate at the nominal level through λ, but does not change the relative order of the test statistics along the genome. As shown in our simulations (model 3–5), when stratification masks the association, genomic control can be quite conservative. Another popular approach to correct for population stratification is structured association [Pritchard et al., 2000] which infers population structure using a set of independent makers. We did not evaluate this method in our simulations due to its computational intensity. Structured association also requires an assumption about the number of underlying subpopulations in the sample. EIGENSTRAT [Price et al., 2006] is an approach for genome-wide association studies based on PCA. It has been shown that the *K*-1 principal components can be related to the solution to the *K*-way clustering solution [Ding and He, 2004]. EIGENSTRAT is less sensitive to the number of components than structured association (if the number is sufficiently large) because of orthogonality of the axes of variation, but the interpretation of the axes is less intuitive.

Our new GSM method tackles the stratification problem by matching at the individual level, without assuming an explicit population structure. Effectively, it treats every sample as a single population and compares it to the most similar counterparts. For samples from clearly distinguished subpopulations, such as the HapMap HCB and JPT populations or the two subpopulations in our simulations, GSM performs almost as well as cluster-based matching or EIGENSTRAT, with little loss of power. In real GWA studies, where sampled individuals may often derive from continuous mixtures of ancestral populations, the individual-based matching in GSM should be more flexible than cluster-based matching. Luca et al. [2008] (GEM) also applied full matching to correct for population stratification, but used a different score calculated from the top eigenvectors from PCA. They showed that outliers may greatly inflate type I errors of association tests using EIGENSTRAT and need to be carefully removed beforehand. The similarity scores in GSM can be used like the GEM scores to identify outliers, but are more intuitive in measuring genetic similarity, compared to the abstract measures from eigenvectors used in GEM. In addition, PCA analysis is very sensitive to the independence of samples, while GSM can actually help to identify related samples through IBS scores. In our Pritzker study example, we found one pair of individuals with large similarity score of 0.85 (S_{IBS}), which strongly suggested a potential first-degree relative. Although the two samples showed strong correlation in their PC scores, they were not identified as outliers by EIGENSTRAT or GEM because their scores did not show strong deviation from the center of the score distributions in the top 10 PCs.

The success of our GSM procedure depends on the accuracy of matching. Incorrectly grouping individuals from different populations could inflate the type I error rate, decrease the power to detect the susceptibility genes, or both. To ensure correct matching, a well-defined similarity measure and a substantial number of markers in which to compute this measure are both important. Our simulations analysis and practical experience, show that similarity measures derived from the distribution of IBS between pairs of individuals, which are simple to calculate and do not require much computing power, provide an effective means of matching individuals. Furthermore, we found that weighting IBS estimates by a function of the marker allele frequencies (S_{freq}) improved the accuracy of matching. Other score metrics also exist and can be easily incorporated into our approach to substitute the IBS-based scores presented. As an experiment, we considered similarity scores based on pairwise IBD estimates calculated using an E-M algorithm, and the average mismatch rates using IBD-based scores were slightly higher than those for IBS-based scores. A weakness of IBD based scores is that they are truncated at zero: when many pairs of individuals are assigned IBD ~0, it becomes difficult to select optimal pairings. Figure 3 demonstrates the relationship between the IBD scores and IBS scores (S_{freq}) computed on the HapMap HCB and JPT samples.

The number of markers used in score calculation is another factor that affects the matching. We prefer to calculate the scores based on a large set of markers (typically including 10,000–100,000 SNPs). However, using too many markers increases the computational load while not necessarily improving the accuracy of matching. In our simulations, 10,000 markers with the smallest *P*-values from one-sided HWE test can correctly match the individuals from closely related populations such as Han Chinese and Japanese, with zero or almost zero mismatch (Table V). In this example, using 30,000 markers worked as well as using 10,000 markers, while using only 1,000 markers led to incorrect grouping of individuals from different populations with up to ~10% mispaired individuals. For samples with subtle differences in genetic ancestry, such as the European American samples in the bipolar data, more markers (50,000–100,000, passing quality control) may help to obtain better matching. Inspecting the genomic control parameter λ on its closeness to the expected value of 1 from different analysis strategies can help to determine the appropriate number of markers for controlling stratification. To select the subset of markers, we usually prefer those with smaller *P*-values from one-sided HWE tests, because they tend to be more informative about population structure. However, we need to be cautious regarding data quality, since markers with high error rates may show strong deviation from HWE and then give incorrect information about the genetic background of sampled individuals. A reasonable compromise is to exclude SNPs with extreme deviations from HWE (say, $P<10^{-6}$) but focus on those with mild deviations (say, $10^{-2}<P<10^{-6}$) to evaluate stratification. GSM does not require that *all* markers should be independent of disease status, since in a typical genome-wide setting the vast majority of markers will meet this criterion and the impact of disease-associated markers on the similarity scores is negligible and can be ignored. Furthermore, since our similarity scores are a function of the mean (weighted) IBS values across a large number of markers, it is also not critical that the assessed SNPs should be independent of each other.

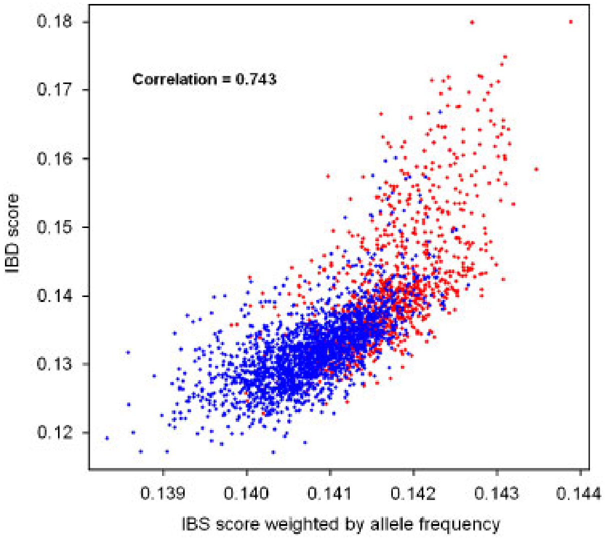

Fig. 3. Similarity scores (calculated from 888,071 SNPs) between each pair of Han Chinese (HCB) and HCB-Japanese (JPT) in HapMap. Red, HCB-HCB pair; blue, HCB-JPT pair.

We chose not to include X-linked markers in our matching scheme to avoid any possible biases due to differences by gender. Given genome-wide association data, the autosomal markers provide ample information for accurate matching.

When there is no population stratification, our simulations showed a small loss of power in GSM due to unnecessary matching. Studies have shown that when the population is indeed homogeneous, random matching by pairs (1:1) can do almost as well as the unmatched test [Chase, 1968]. Additional power may be lost when the matching is not balanced, so that multiple controls are compared to a single case subject or multiple cases are compared to a single control (i.e., 1:*m* or *m*:1 when $m>1$). However, when stratification is present, larger values of *m* are preferred to decrease the chance of matching errors. It is then a trade-off of efficiency and bias that we need to consider in practice. In our GSM method, the objective function (*T*) we choose for optimal matching favors smaller groups, minimizing loss of efficiency. Although the original optimal matching [Rosenbaum, 1991] is unconstrained ($m=\infty$) so that all controls are allowed to be matched to a single case or all cases to a single control, Hansen [2004] showed that the matching with restriction on *m* can reduce the variance of estimated parameters with little increase in bias, and suggested a linear search for good values of *m* that are as close to 1 as possible. In our simulations, a large proportion of the matched sets are 1:1 matches even when the proportions of the two populations in cases and controls are not equal, and the average size of matched sets does not vary much for different values of the upper bound of *m*. For example, for simulated setting 3, the average matched set size is 2.44 and 2.47 when the upper limits of *m* are set as 2 and 5, respectively.

Although the full matching scheme is flexible, cases (or controls) from a population without a corresponding partner among the controls (or cases) will decrease power and may lead to spurious association if matching is forced. Further, 1:1 matching is more efficient than *m*:1 for $m>1$. Therefore, we still strongly encourage careful sample selection during the study design. Skol et al. [2005]

showed that the self-reported ethnicity can be a good predictor for population structure, consistent with our results based on the NIMH case and control samples alone.

In summary, we propose a new framework to match case and control samples by their genetic similarity and adjust for the underlying population substructure. Our GSM method is specifically designed to use the full information provided by the large number of genotypes in genome-wide association studies or large-scale candidate gene studies. Our method can correctly control the false positives, while maintaining considerable power to detect the disease-marker association. Our individual-based matching scheme can reflect the continuous mixing of ancestral populations. By comparing each case to one or more controls sharing the most genetic backgrounds, we hope our method may increase the chance to identify the genetic variants that influence disease risk. Our GSM software is available freely with C++ source code at http://www.sph.umich.edu/csg/liang/gsm/. The package allows the users to automatically calculate matching score matrices, conduct full matching with a range of parameter choices, and carry out association analyses. We expect our method will aid analyses of large-scale genome-wide association studies.

ACKNOWLEDGMENTS

We thank the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. We thank NIMH Human Genetics Initiative for providing the case and control samples, and the Illumina database and its contributors for providing the additional control samples. This research was supported by grants HG000376 (M.B.) and HG02651 and HL084729 (G.R.A.) from the National Institutes of Health.

REFERENCES

Bertsekas DP, Tseng P. 1994. RELAX-IV: a faster version of the RELAX code for solving minimum cost flow problems. Technical Report LIDS-P-2276, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Chase GR. 1968. On the efficiency of matched pairs in Bernouilli trials. Biometrika 55:365–369.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Ding C, He X. 2004. *K*-means clustering via principal component analysis. Proceedings of the International Conference on Machine Learning (ICML 2004), Banff, Canada. p 225–232.

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421.

Epstein MP, Allen AS, Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 80:921–930.

Frangioni A, Manca A. 2006. A computational study of cost reoptimization for min cost flow problem. INFORMS J Comput 18:61–70.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388–393.

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. 2006. Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 30:277–289.

Guan W, Liang L, Boehnke M, Abecasis GR. 2005. Matching cases and controls using genotype data from a whole genome association study. ASHG 2005 Annual Meeting, Salt Lake City, Utah, #2395 (poster).

Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. J Am Stat Assoc 99:609–618.

Hansen BB, Klopfer SO. 2006. Optimal full matching and related designs via network flows. J Comput Graph Stat 15:609–627.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1–44.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007. A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Genet 81:895–905.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.

Li CC. 1972. Population subdivision with respect to multiple alleles. Ann Hum Genet 33:23–29.

Liang L, Zöllner S, Abecasis GR. 2007. GENOME: a rapid coalescent-based whole genome simulator. Bioinfomatics 23:1565–1567.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453–463.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685–693.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. Am J Hum Genet 67:170–181.

Rosenbaum PR. 1991. A characterization of optimal designs for observational studies. J R Stat Soc Ser B 53:597–610.

Rosenbaum PR. 2002. Observational Studies, 2nd edition. New York: Springer.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345.

Skol AD, Xiao R, Boehnke M, Veterans Affairs Cooperative Study 366 Investigators. 2005. An algorithm to construct genetically similar subsets of families with the use of self-reported ethnicity information. Am J Hum Genet 77:346–354.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–513.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy–Weinberg equilibrium. Am J Hum Genet 76:887–893.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341.

Zhang S, Kidd KK, Zhao H. 2002. Detecting genetic association in case-control studies using similarity-based association tests. Stat Sin 12:337–359.

APPENDIX A

In a minimum cost flow (MCF) problem, we define a directed graph consisting of nodes, $i \in \mathcal{N}$, and arcs connecting the nodes, $(i,j) \in \mathcal{A}$. For each arc (i,j), an integer a_{ij} denotes the cost and a positive integer c_{ij} the capacity. For each node i, an integer s_i denotes the exogenous supply. A solution of the MCF problem is a set of arc flows x_{ij} that minimizes

$$\sum_{(i,j)\in\mathcal{A}} a_{ij}x_{ij}$$

subject to the constraints on capacity:

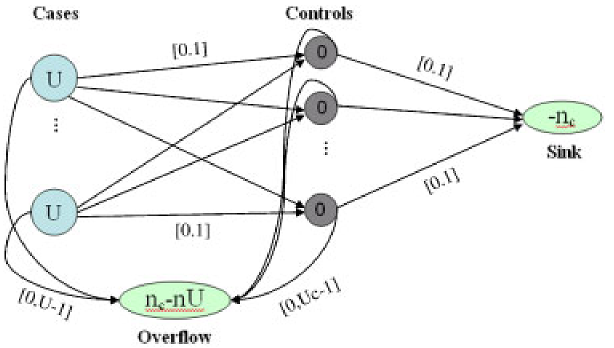

Fig. 4. Solve optimal full matching problem as a minimum cost flow (MCF) problem. *U* denotes the maximal number of controls each case can match, *U*c the maximal number of cases each control can match, *n*c the number of controls to match, and *n* the total number of cases and controls.

$$\sum_{\{j|(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j|(j,i)\in\mathcal{A}\}} x_{ji} = s_i \quad \text{for all } i \in \mathcal{N} \quad 0 \le x_{ij} \le c_{ij}$$
$$\text{for all } (i,j) \in \mathcal{A}.$$

It is easy to see the equivalence between the MCF and the optimal matching (Fig. 4). The nodes in a directed graph correspond to the cases and controls, a_{ij} is the dissimilarity measure between i and j, and the capacity of the flow, c_{ij}, is 1 between case and control nodes, and 0 between two cases or two controls. The optimal solution of the MCF problem is equivalent to an optimal matching. The nodes connected by arcs with non-zero flow are assigned to the same matched set.

In full matching, the numbers of case-control pairs vary across matched sets, so the supply of nodes (s_i) cannot be predetermined. To deal with this complication, we include an ''overflow'' node to the graph to balance the flows from or to the case or control nodes. Parameters U and U_c control the maximum flows going to ''overflow'' from each node, which correspond to the maximum number of cases or controls allowed in each matched set, i.e., the upper limit of m in 1:m or m:1 match. For each case node, there are m connected control nodes and U-m arcs connecting it to ''overflow''; for each control node, there are m connected case nodes and m arcs connecting to ''overflow.'' The cost for arcs entering ''overflow'' is set as 0, so these extra arcs do not affect the total cost. Similarly, another node, ''sink,'' may also be added to control the total number of controls to be matched, and the cost for arcs entering ''sink'' is also 0 (Hansen and Klopfer, 2006).

The translation is demonstrated in Figure 4. The MCF problem is then solved by iteratively updating a dual cost vector and the flow vector x (Bertsekas and Tseng, 1994; Frangioni and Manca, 2006).