Am. J. Hum. Genet. 74:317–325, 2004

Matching Strategies for Genetic Association Studies in Structured Populations

David A. Hinds,[1] Renee P. Stokowski,[1] Nila Patil,[1] Karel Konvicka,[1] David Kershenobich,[2] David R. Cox,[1] and Dennis G. Ballinger[1]

[1]Perlegen Sciences, Mountain View, CA, and [2]Departmento de Gastroenterología, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City

Association studies in populations that are genetically heterogeneous can yield large numbers of spurious associations if population subgroups are unequally represented among cases and controls. This problem is particularly acute for studies involving pooled genotyping of very large numbers of single-nucleotide–polymorphism (SNP) markers, because most methods for analysis of association in structured populations require individual genotyping data. In this study, we present several strategies for matching case and control pools to have similar genetic compositions, based on ancestry information inferred from genotype data for ~300 SNPs tiled on an oligonucleotide-based genotyping array. We also discuss methods for measuring the impact of population stratification on an association study. Results for an admixed population and a phenotype strongly confounded with ancestry show that these simple matching strategies can effectively mitigate the impact of population stratification.

Introduction

Genomewide association studies provide a powerful approach to implicate DNA variants (and, by extension, the genomic regions they represent) in the predisposition to complex diseases and in the genetic underpinnings of drug efficacy and adverse reactions. The success of these studies relies on the accurate measurement or estimation of allele-frequency differences between case and control subjects. When searching for small genetic effects in large association studies, systematic differences in ancestry between the cases and controls are likely to produce many statistically significant but spurious associations (e.g., Knowler et al. 1988; Lander and Schork 1994). Such differences are expected to be found when genetically distinct population subgroups have a different prevalence of the target phenotype.

The use of family-based association study designs mitigates the impact of systematic ancestry differences (population stratification) but can lead to an increased burden in the recruitment of subjects and in genotyping (Cardon and Palmer 2003). Self-reported ancestry is also useful in matching case and control subjects to reduce the prevalence of spurious associations. Population structure can be empirically determined by individually genotyping all potential cases and controls across a set of unlinked marker loci (Pritchard and Rosenberg 1999). When individual genotypes are known, analysis methods can correct the association test statistic for unmatched groups by use of the inferred population structure (Pritchard et al. 2000*b*; Reich and Goldstein 2001; Satten et al. 2001; Thornsberry et al. 2001; Hoggart et al. 2003).

In association studies using DNA pooled from many individuals, significant causal disease (or pharmacogenetic) associations would be indistinguishable from associations due to ancestry differences between cases and controls. Thus, genetic-ancestry matching prior to DNA pooling is essential. By use of inferred population-structure data, DNA pools can be constructed that are matched to have similar genetic composition, to minimize the likelihood of spurious associations due to population stratification. Allele-frequency estimates in the matched DNA pools should then give a more reliable indication of causal disease association. See the work of Sham et al. (2002) for a recent review of DNA pooling methodologies and implications for association studies.

In genomewide association studies, it is necessary to test at least hundreds of thousands of SNP markers because of the generally limited extent of linkage disequilibrium in the human genome (Risch and Merikangas 1996; Kruglyak 1999; Risch 2000; Patil et al. 2001). We are currently testing >1.5 million SNP markers in association studies, using pooled genotyping with multiple measurements of allele frequency in each of two pools as an efficient screen to enrich for SNPs with significant allele-frequency differences. The SNPs with

Received August 29, 2003; accepted for publication November 26, 2003; electronically published January 21, 2004.

Address for correspondence and reprints: Dr. David Hinds, Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 94043. E-mail: David_Hinds@perlegen.com

0002-9297/2004/7402-0013$15.00

the greatest apparent allele-frequency differences in the pooled data are then selected for individual genotyping. The pooled genotyping step reduces the number of SNPs that must be individually genotyped to confirm allele-frequency differences between case and control groups. In this context, spurious associations due to population structure force us either to examine more SNPs by individual genotyping or, if that is impractical, to sacrifice power to detect causal associations.

In this study, we describe the use of unlinked SNP markers to detect and correct for population stratification in case and control subjects in an admixed population prior to pooled genotyping for association testing. Using a phenotype that is strongly confounded with ancestry, we show that several strategies for matching case and control groups are successful at eliminating significant stratification. We also discuss methods for measuring the impact of stratification on a pooled genotyping experiment.

Methods

Subject Collection

Subjects were chronic alcoholics, some with alcoholic liver disease, recruited in Mexico City under full informed consent. The international institutional review board of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ), which is registered with the Office of Human Research Protection, approved the human patient sample-collection protocol. Subjects were measured for height in cm at the time of blood-sample collection. The three self-reported ethnicities in this population were "Caucasian," those of primarily Spanish European ancestry; "Otomi" Indians, from the Pachuca region in Mexico; and "Mestizo," a mix of Spanish European and Mexican Indian ancestry. A total of 824 Mestizo males were examined to determine the distribution of height. The definitions of "tall" and "short" were chosen to include the upper and lower 25% of the observed distribution. This yielded a minimum height of 174 cm for the "tall" group and a maximum height of 162 cm for the "short" group.

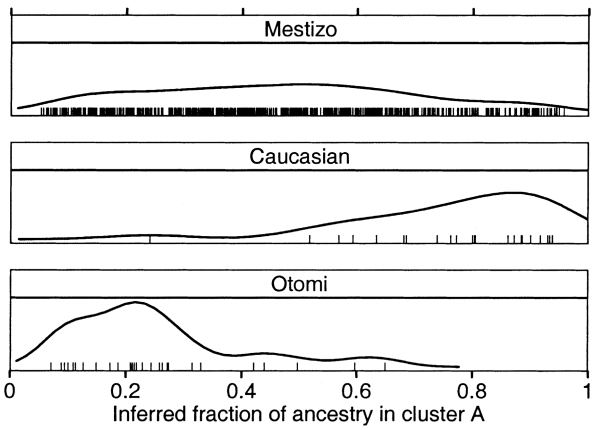

Figure 1 Distribution of ancestry for self-reported population subgroups. Density distributions for the inferred fraction of subjects with cluster A ancestry are shown for 655 Mestizo, 23 Caucasian, and 29 Otomi Indian subjects. Each tick mark represents the fractional ancestry of an individual subject.

SNP Selection

From a genomewide collection of SNPs discovered by Perlegen Sciences in a globally diverse panel of individuals (Patil et al. 2001), we selected a set of 312 that were roughly equally spaced across the autosomes and were expected to behave well in oligonucleotide array–based genotyping. SNPs were selected to be at least 150 bp from the nearest common repetitive element, as identified by the RepeatMasker 2 program (available on the RepeatMasker Web site), and the 25-bp sequence containing the SNP ($\pm$ 12 bases of context) was required to be unique in the human genome, according to then-current National Center for Biotechnology Information (NCBI) Build 29 (available on the NCBI Web site). We also required that in Perlegen's previously collected SNP discovery data, the SNPs have a high rate of high-confidence genotype calls and an allele frequency close to 0.5. A combination of these quality metrics was used to numerically score each candidate SNP. We then selected the highest-scoring candidates from a series of 2-Mb windows spaced at 9-Mb intervals across each NCBI Build 29 chromosome.

Primer Design

PCR primer pairs for each SNP were selected using the program Oligo, version 6.57 (Molecular Biology Insights). We selected primers having a T_m of 59°C–66°C, a length of 18–22 bases, a PCR product size of 50–200 bases, and 3′-end ΔG of between -5.5 and -9.8 kcal/mol. We also required that each primer be at least 5 bases from its target SNP. Primer sequences containing repetitive sequences, as determined by the RepeatMasker 2 program, were excluded. Only primer sequences determined to be unique ($P < 10^{-4}$) in the genome (NCBI Build 29) by use of the BLAST program (available on the NCBI BLAST Web site) (Altschul et al. 1990) were selected.

Genotyping Oligonucleotide Array Design

Genotyping arrays of 25-bp oligonucleotides were designed as four sets of 20 features (80 features per SNP), corresponding to forward and reverse strand tilings for sequences complementary to each of two SNP alleles. A set of 20 features consisted of five sets of 4 features where the location of the SNP within the oligonucleotide varied

Table 1

Quality-Control Checks for SNP Genotyping Results

Data-Quality-Filter Criterion	No. of SNPs Passing	% Passing
Pass rate >80%	309	99%
Three genotype clusters identified	308	99%
<20 ambiguous calls	305	98%
$P > .00001$ for Hardy-Weinberg equilibrium	303	97%
Maximum cluster width	282	90%
All criteria	275	88%

from position 11 to position 15. A set of 4 features consisted of sequences where A, C, T, or G was substituted at position 13. Thus, each set of four features provided one perfect match to the sequence of the corresponding SNP allele and three features with a single-base mismatch for that allele. Mismatch probes were used to measure background and, by comparison with the signal for the perfect match probes, to detect the presence or absence of a specific PCR product in a sample. Light-directed chemical synthesis of the appropriate oligonucleotides was carried out by Affymetrix (Fodor et al. 1991).

Hybridization Sample Preparation

For analysis of the 312 stratification SNPs, DNA was amplified by PCR in 12-μl volume containing 13 primer pairs at 0.4 mM of each primer, 10 ng of individual genomic DNA, 2 U Titanium *Taq* (Clontech), 0.5 mM deoxynucleotide triphosphates, 10 mM Tris-HCl (pH 9.1), 3 mM $MgCl_2$, and additives. Thermocycling was performed on a 9700 cycler (Perkin-Elmer), with initial denaturation at 96°C for 5 min, followed by 10 cycles of 96°C for 30 s, 58°C minus 0.5°C/cycle for 30 s, 65°C for 1 min, then 40 cycles of 96°C for 10 s, 53°C for 30 s, and 65°C for 60 s, and, finally, an extension at 65°C for 7 min. PCR products were pooled together and labeled with 0.7 μM biotin-16-ddUTP/dUTP (Roche) with 25 units of terminal deoxynucleotidyl transferase (Roche), by incubating at 37°C for 90 min, after which the reaction was stopped by heat-inactivation at 99°C for 10 min.

Hybridization of Samples to High-Density Oligonucleotide Arrays

Labeled DNA samples were incubated in hybridization buffer (3 M tetramethylammonium chloride, 10 mM Tris-HCl [pH 7.8], 0.01% Triton X-100, 100 μg/ml herring sperm DNA, and 50 pM control oligomer) at 99°C for 10 min and hybridized to a chip overnight at 50°C on a rotisserie at 25 rpm. Chips were washed twice in 1 × MES buffer (0.1 M 2-[N-morpholine]ethane sulfonic acid [pH 6.7], 1 M NaCl, and 0.01% Triton X-100), and incubated with 5 μg/ml streptavidin (Sigma-Aldrich) and 2.5 mg/ml acetylated bovine serum albumin (Sigma-Aldrich) in 1 × MES for 15 min on a rotisserie at room temperature (RT). After two washes with 1 × MES at 35°C, chips were incubated with antibody solution (1.25 μg/ml biotinylated antistreptavidin antibody [Vector Laboratories] and 2.5 mg/ml BSA in 1 × MES) for 15 min on a rotisserie at RT, followed by another two washes with 1 × MES at 35°C. Then, chips were stained with 1 μg/ml streptavidin-Cy-chrome conjugate (Molecular Probes) and 2.5 mg/ml BSA for 15 min on a rotisserie at RT, followed by two washes with 1 × MES at 35°C. Chips were incubated for 30 min at 37°C in 0.2 × SSPET (30 mM NaCl, 2 mM NaH_2 PO_4, 0.2 mM EDTA [pH 7.4], 0.01% Triton X-100), followed by a wash with 1 × MES at RT. Hybridization of the labeled sample to the chip was detected using a confocal laser scanner (Perlegen) (Patil et al. 2001).

SNP Genotyping

For each SNP, we measured ratios of the mean intensity of perfect-match features for one allele to the sum of mean intensities for both alleles. In principle, these ratios should take on values near 1.0, 0.5, or 0.0 for AA, AB, or BB genotypes. We discarded data if, for both alleles, <9 out of 10 perfect-match features were brighter than their corresponding mismatch features. We used an expectation-maximization algorithm and a normal mixture model to assign intensity ratios to clusters.

For the stratification analyses, we only used data for SNPs that showed consistently good genotyping results (table 1). We excluded SNPs that had a pass rate of <80% on the basis of the perfect-match/mismatch comparison. We also excluded SNPs for which fewer than three genotype clusters could be identified, as well as those that had >20 ambiguous cluster assignments. Many SNPs showed moderate departures from Hardy-Weinberg equilibrium, which would be expected in a heterogeneous population. We excluded only those SNPs showing extreme deviations that could be traced back to convergence failures of the clustering algorithm. For the 275 SNPs passing these criteria, the overall call rate was 98.4%. In a set of 24 individuals genotyped in triplicate for these SNPs, we had a concordance of 99.8%. The 275 SNPs and all individual genotype data used in this study have been submitted to dbSNP (ss12673803–ss12674077) (available on the dbSNP Web site). SNP

positions in NCBI Build 33 are also shown in table A (online only).

Statistical Analysis

We used the *structure* program (Pritchard et al. 2000*a*) to identify population subgroups and infer admixture information from SNP genotype data. All runs were 100,000 cycles, after a 20,000-cycle burn-in period. We selected a model with admixture and with correlated allele frequencies; we used the defaults for other settings. We did not use prior information about population membership to direct the clustering. Without this information, the *structure* program cannot distinguish between solutions with permuted cluster labels; therefore, we manually assigned labels to clusters, for consistency across multiple analyses. Genetic distances (F_{ST}) were calculated from *structure*'s allele-frequency estimates, as in the study by Weir (1996). False-discovery rates were calculated using Q-VALUE (available on the Q-VALUE Software Web site) (Storey and Tibshirani 2003). All other statistical analyses were performed with the R package (available on the R Project Web site) (Ihaka and Gentleman 1996).

Results

Assessment of Population Structure

A total of 707 individuals recruited in Mexico City were selected for genotyping. The majority of subjects (655) were of Mestizo ("mixed") ancestry; small numbers of individuals of self-reported Caucasian (23) and Otomi Indian (29) ancestry were also included. Using high-density oligonucleotide arrays, we genotyped these subjects for 312 uniformly spaced, unlinked SNPs. Of the 312 markers, 275 yielded high-quality genotype data. Many of the SNPs showed larger-than-expected allele-frequency differences between the three subpopulations, measured as an excess of small P values in χ^2 tests (table 2). Controlling for false-discovery rate (Storey and Tibshirani 2003), we also counted SNPs having q values < 0.05 and found many significant associations. The q value method accounts for multiple testing, and it indicates the number of SNPs with significant associations such that, on average, only 5% will be false positives.

Table 2

Association Test Results for Population Subgroups with 275 SNPs

	Number of SNPs with χ^2 Test Statistics				
	$P < .0001$	$P < .001$	$P < .01$	$P < .1$	$q < .05$
Expected	0	0	2.75	27.5	0
Caucasian—Mestizo	2	5	23	85	8
Otomi—Mestizo	0	1	15	50	0
Otomi—Caucasian	3	14	34	105	32

Figure 2 Distribution of ancestry versus height categories. Density distributions for the inferred fraction of subjects with cluster A ancestry are shown for 164 short and 166 tall subjects. Each tick mark represents the fractional ancestry of an individual subject.

We analyzed this genotype data for population structure using the *structure* program (Pritchard et al. 2000*a*; available on the Pritchard Lab Web site). This is a model-based method for identifying subpopulations in which, within each subpopulation, all markers are in Hardy-Weinberg and linkage equilibrium. The analysis supported the presence of two genetically distinct population clusters, one of mostly European ancestry ("cluster A"), and one of mostly Indian ancestry ("cluster B"). The estimated cluster-membership proportions for self-reported Caucasian and Otomi Indian samples are well separated; Mestizo samples are uniformly distributed across nearly the full range of values (fig. 1). There was no strong evidence for models with more than two population clusters. On the basis of their estimated allele frequencies, we determined a genetic distance of $F_{ST} = 0.14$ between the two clusters. Phenotype information and cluster-membership proportions for each sample are reported in table B (online only).

The admixture model used in the *structure* program assumes a unimodal distribution of individual admixture proportions. However, we found that our inclusion of small numbers of Caucasian and Otomi samples in the analysis did not significantly perturb the admixture estimates for the Mestizo samples. A separate analysis of just the Mestizo samples, which might be expected to better fit the unimodal admixture model, yielded admixture proportions that had a correlation of 0.9994 with the full analysis (data not shown). Thus, this analysis seems to be robust against some limited misspecification of the admixture model.

Association of Ancestry with Height

We compared the inferred ancestry information for individuals selected to represent the tallest and shortest 25% of male Mestizo subjects. Of the samples that were

genotyped, we identified 164 short and 166 tall individuals. Height is strongly correlated with the inferred proportion of cluster A ancestry (fig. 2), and many spurious allele-frequency differences occur solely as a result of differences in ancestry between the tall and short groups (table 4, all samples). This is an extremely stratified population, and there are multiple SNPs with χ^2-test P values of $<10^{-8}$. This level of significance would exceed genomewide significance thresholds for 1 million independent SNP association tests with conservative adjustment for multiple testing, clearly a problem if these groups were to be used in the type of genomewide association study described above.

Matching Based on Average Ancestry Estimates

We composed new groups using subsets of the tall and short individuals, so the groups would have the same average proportions of ancestry in clusters A and B, while retaining as many samples as possible. This involved removing tall samples with the highest proportions of cluster A ancestry and short samples with the lowest proportions of cluster A ancestry. We were able to retain 98 tall samples and 98 short samples with this matching strategy. Ancestry proportions before and after matching are shown in table 3. For a direct comparison, 98 samples were also selected at random from the lists of tall and short samples. The random and matched groups were tested for significant allele-frequency differences (table 4, random and matched subsets). Matching removed most evidence for population structure. An overall test for stratification that was based on the sum of χ^2 statistics (Pritchard and Rosenberg 1999) for the matched set gave a P value of ~.005, versus $\sim 10^{-71}$ for the randomly selected set. The distribution of P values for the 275 SNPs is more nearly uniform for the matched groups (fig. 3), and no markers showed significant association after controlling the false-discovery rate.

In the previous analysis, the SNPs used to test for associations were the same ones used for the stratification analysis. Although the stratification analysis is blind to the phenotype, in principle, this analysis could underestimate the residual population structure expected for other SNPs not included in the stratification analysis. To address this, we split the 275 SNPs into five random subsets of 55. For each subset, we performed a stratification analysis of the other 220 SNPs, matched tall and short groups on the basis of that analysis, and then tested for association in the 20% that had been left out. Then we combined results for all the subsets, yielding a test result for each SNP stratified by use of what was, for that SNP, an independent set of data. Results (table 4, leave-out-20% data set) were essentially the same as for matching on all 275 SNPs, and there were no significant associations.

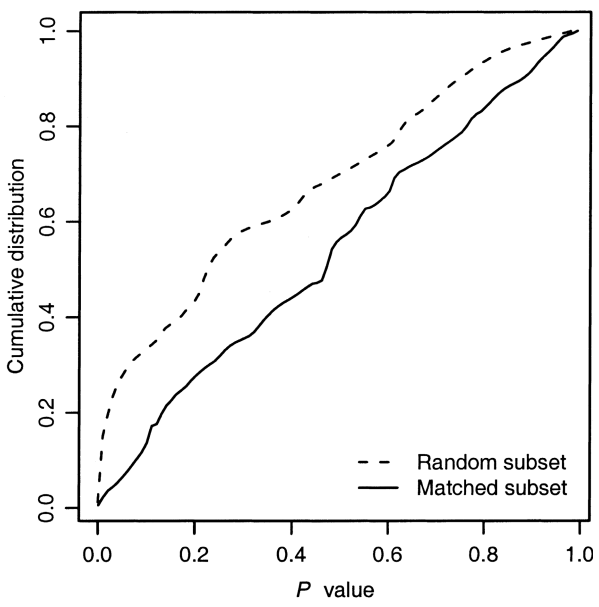

Figure 3 Cumulative distribution of P values for 275 SNPs, for the random and ancestry-matched subsets of tall and short subjects. In the absence of population structure, the P values should be uniformly distributed, and their cumulative distribution should be a straight line from (0,0) to (1,1). The random subset shows an excess of small P values, whereas the matched subset has a nearly uniform distribution.

Matching Based on an Ancestry-Adjusted Phenotype

An alternative approach to eliminating stratification for a quantitative trait is to define groups on the basis of a phenotype that has been adjusted to remove effects of ancestry differences. We performed a linear regression

Table 3

Average Proportion of Ancestry in Cluster A, for Tall and Short Groups

	No. of Subjects in		Proportion of Ancestry in Cluster A in	
Data Set	Tall Group	Short Group	Tall Group	Short Group
All samples	166	164	.62	.36
Matched subset	98	98	.48	.48

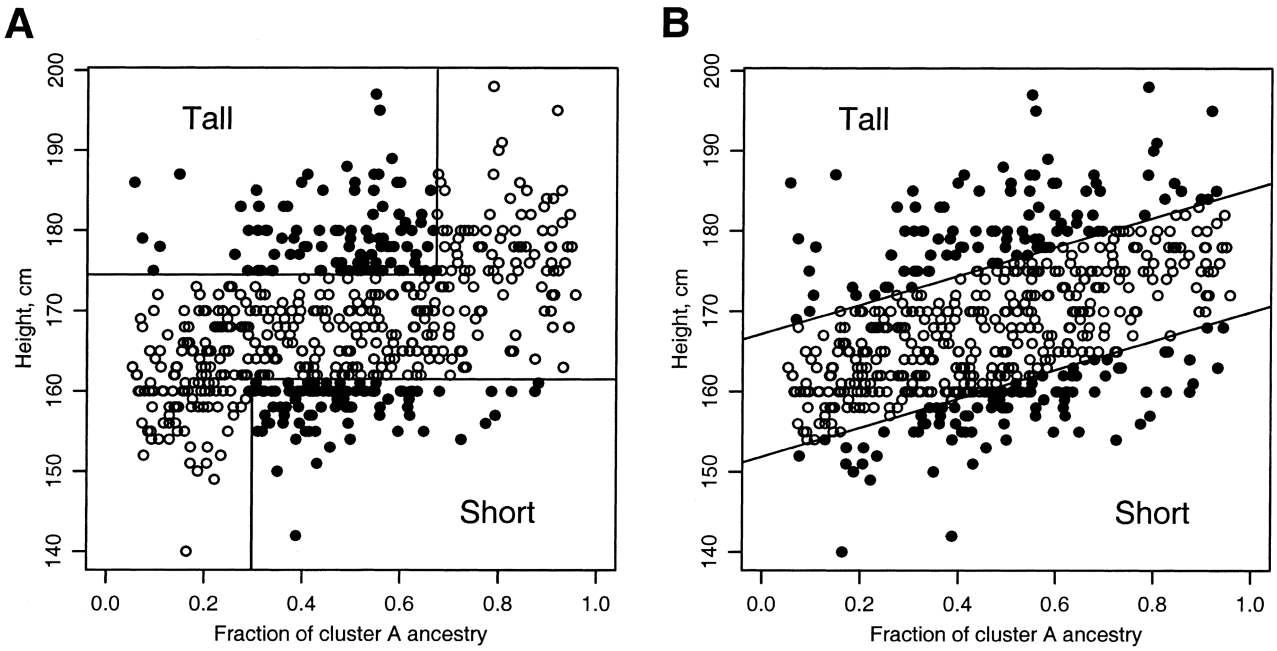

Figure 4 Comparison of a matching strategy with independently determined cutoffs for height and ancestry (*A*) and a strategy based on a linear regression of height against ancestry (*B*). The samples retained from tall and short subjects by use of each method are shown as blackened circles, and excluded samples are shown as unblackened circles. The regression method results in inclusion of the tallest and shortest individuals within any narrow window of ancestry values.

of height against the inferred fraction of cluster A ancestry for the male Mestizo subjects in our study and determined that a 10% increase in cluster A ancestry corresponded, on average, to a 1.8-cm increase in height. We adjusted height by subtracting out this contribution, and we selected the tallest and shortest 98 individuals on the basis of the adjusted phenotype. We did not see any significant associations using these groups (table 4, linear adjusted).

In principle, adjusting for ancestry should yield a cleaner phenotype and a more powerful study design than the simple strategy of matching the mean ancestry of case and control groups. Comparing the distributions of height and inferred ancestry for the two designs (fig. 4), the regression design includes fewer individuals with relatively mild ancestry-adjusted phenotypes and intermediate ancestry coefficients, and more individuals with extreme ancestry-adjusted phenotypes and ancestry coefficients. The regression design may be more challenging to implement, however, if it requires collecting genotype data for additional individuals to accurately determine the relationship between phenotype and ancestry.

Effects of Population Structure on Pooled Genotyping

In many if not most association studies, if the target population is relatively homogeneous, or if there is little confounding between the target phenotype and ancestry, then careful pool matching may not be necessary (e.g., Ardlie et al. 2002). Thus, it is useful to have a way of quantifying the practical impact of population structure on an association study, to decide when corrective action is needed. Significance tests are not appropriate for this purpose because they do not directly measure the magnitude of an effect. One approach is to model population structure as one of various sources of error that lead to an increase in the false-positive rate. If the effect of population structure is determined to be small compared with other known sources of experimental error, then correcting for it will have limited benefit.

We examined the behavior of the sum of χ^2 statistics for association tests with data from the tall and short groups matched for average ancestry, as various amounts of random noise were added to allele frequencies in the two groups. Genotypes for each SNP were first permuted to eliminate any residual disequilibrium, so we essentially only preserved overall SNP allele frequencies from the original data. The allele frequencies for each pool were then perturbed by a normally distributed error term, with standard deviation specified in units of allele frequency (fig. 5). The sum statistics for the unpermuted random and matched groups (table 5)—that is, 928 and 338—are comparable to permuted data with additional experimental error of ~5% and ~1%, respectively. Additional error on the order of 1% seems tolerable for

currently available pooled genotyping technologies, which generally cannot determine allele frequencies with better accuracy than that (Sham et al. 2002). This approach could be combined with estimates of experimental variance components (Barratt et al. 2002) to produce more realistic end-to-end power estimates for pooled genotyping study designs.

Genomic control (Devlin and Roeder 1999) provides another approach to estimating the magnitude of the effect of population structure in an association study. In this approach, rather than modeling structure in a population, its effects are measured by the inflation of test statistics for markers that, in aggregate, should not show evidence for association. We estimated the variance-inflation factors (λ) due to population structure for each set of tall and short groups by use of this approach. One interpretation of the variance inflation is as a reduction in effective sample size (ESS), which we estimate here as (N / λ), where N is the original sample size (table 5). Genomic control would effectively maintain a desired type I error rate in the presence of population structure in this example; however, it does so at a substantial cost in the ESS and, hence, power to detect causal associations. Our results show that matching to mitigate the impact of population structure can substantially boost the ESS, despite the reduction in raw sample count.

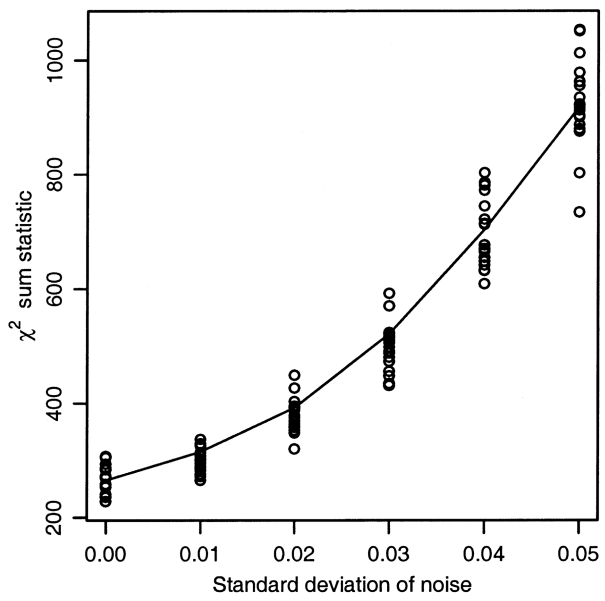

Figure 5 Effect of simulated experimental error on an overall population-structure test statistic. We simulated the effect of experimental error by adding normally distributed noise to allele-frequency estimates in permuted copies of the genotype data for the matched tall and short groups. The overall test statistic is the sum of resulting χ^2 statistics for the 275 individual SNPs; this is expected to follow a χ^2 distribution, with 275 df. We show results for 20 separate permutations for each value of the noise parameter.

Discussion

Our results indicate that relatively simple matching strategies can effectively control for population stratification in case-control association studies, for a phenotype with a very large ancestry effect in an admixed population. The genotyping can be efficiently implemented in the laboratory in a high-throughput setting, with a single generic SNP genotyping array carrying around 300 uniformly distributed SNPs that are chosen without regard to their allele frequencies in specific target populations. We have now processed many thousands of these arrays.

Although we chose to use the *structure* program to infer admixture proportions, other methods are available, including the ADMIXMAP program (available on the Genetic Epidemiology Group Web site) (McKeigue et al. 2000; Hoggart et al. 2003), which may offer significant benefits in some situations. The admixture model in *structure* suffers from a theoretical deficiency (Pritchard et al. 2000*a*; Hoggart et al. 2003), in that it does not permit specification of prior allele-frequency information for the ancestral populations and thus cannot disambiguate between symmetric modes that differ only in the labels assigned to clusters. Also, interpretation of the admixture coefficients relies on the sampler only exploring one of these symmetric modes. In our analysis, we verified that individual *structure* runs consistently settled in one (randomly selected) mode, and we could easily determine consistent cluster labels when comparing results across multiple runs. The matching strategies we describe are also invariant under permutations of the cluster labels. Still, it is possible that the *structure* sampler may have more trouble in situations with more clusters or less clearly separated ones.

In the context of a pooled genotyping screen, absolute control of population structure is probably not required in many cases. It is probably only necessary to ensure that the incremental increase in variance due to population differences between case and control pools is small compared with other sources of variance in the genotyping experiment. In an association study design consisting of an initial screen of many markers by

Table 4

Association Test Results for Height in 275 SNPs

	Number of SNPs with χ^2 Test Statistics				
Data Set	$P < .0001$	$P < .001$	$P < .01$	$P < .1$	$q < .05$
Expected	0	0	2.75	27.5	0
All samples	22	38	69	126	94
Random subset	10	20	44	106	62
Matched subset	0	0	7	35	0
Leave out 20%	0	0	6	39	0
Linear adjusted	0	0	4	44	0

Table 5

Overall Measures of Population Structure for Height Pools

Data Set	χ^2 Sum	*P* Value	λ^a	ESS
All samples	1,380	2×10^{-146}	4.9	34
Random subset	928	8×10^{-72}	3.6	27
Matched subset	338	5×10^{-3}	1.1	89
Leave out 20%	345	2×10^{-3}	1.4	70
Linear adjusted	313	6×10^{-2}	1.3	75

[a] Variance-inflation factor, calculated as follows: median(χ^2)/0.456.

pooled genotyping followed by individual genotyping of candidates, there should be more tolerance for spurious associations in the pooled step. In these cases, a test for population structure on a representative subset of cases and controls may be sufficient to place bounds on the impact of population stratification on the entire study, thus avoiding unnecessary recruitment or individual genotyping effort.

A complete association study would consist of three phases. First, some or all samples would be individually genotyped to ascertain their population structure using our array of ~300 SNPs. On the basis of those results, and constrained by the form of the phenotype and its ascertainment method, a strategy for mitigating population structure would be selected and validated using the available genotype data. Both of our matching strategies require genotyping some individuals who will end up being excluded from the matched case and control pools. The second phase would consist of pooled genotyping of many SNPs in replicate experiments. In a third phase, candidate SNPs would be selected for individual genotyping on the basis of the pooled data. Samples originally excluded from the pools could be genotyped at this point and could be analyzed using one of the structured association approaches. Genomic control could also be used to adjust significance tests for any residual population structure left in the matched pools.

The matching strategies we discuss here were designed for whole-genome association studies for which we required that a solution could not increase the experimental effort required at the pooled genotyping stage. This constraint (a practical, economic one) severely limited the range of solutions that we could consider. Another approach to controlling for population structure would be to perform a stratified analysis of subpools composed of individuals of similar ancestry. For experimental designs permitting many replicates, this may be a useful strategy for discrete traits that cannot be adjusted to remove ancestry effects. Such a design would allow all individuals to be included in the pooled analysis; however, strata with very unbalanced representation of the trait values would have somewhat lower informativeness for equal experimental effort. The number of strata required to account for most of the variance in ancestry would multiply the experimental effort required for allele-frequency determination, since this would be orthogonal to any replication required to characterize experimental variance.

The strategies we describe can be extended to more complex structured populations. For either admixed populations or populations composed of several unadmixed groups, our approach would be either to match the average genetic contribution of each empirically identified cluster in the case and control groups by excluding samples, or to use multivariate regression to determine an ancestry-adjusted phenotype for each individual on the basis of the individual's inferred cluster-membership proportions. In the absence of admixture, a multiethnic pooled study would be most sensitive for detecting loci that account for phenotypic variation in all of the included populations; such a study would be insensitive for loci accounting for fixed differences between populations.

Admixed populations are attractive targets for association studies because these groups should show more linkage disequilibrium over larger physical distances (Chakraborty and Weiss 1988). If the admixture is between populations with significantly different genetic predispositions to a target phenotype, then heritability of a trait in the admixed population may also be higher than in the more homogeneous ancestral populations. Although linkage-based admixture mapping (McKeigue 1998) can be a more efficient approach for identifying loci that specifically explain phenotypic variance between populations, an association study in an admixed population has the ability to detect loci that explain variance either between or within populations. Pooled allele-frequency differences would not distinguish within- from between-population associations, but these could be resolved later by modeling ancestry effects at associated loci by use of individual genotyping data. The groups used in this study are small, and larger sample sizes would be required for a whole-genome association study of a complex multigenic phenotype. The impact of stratification would be correspondingly larger for more realistic study designs, because although sampling variation in allele frequencies becomes smaller for larger pool sizes, the variance due to population stratification does not. Careful management of population structure is likely to be an important component of future whole-genome association studies.

Acknowledgments

We wish to acknowledge Dr. Raul Bernal Reyes (Instituto Mexicano del Seguro Social, Pachuca), Dr. Armando Diaz Belmont (Hospital General de México), and A. Christian Perez Pruna and Marta Garcia Sandoval (Instituto Nacional de Nu-

trición Salvador Zubirán), for their invaluable efforts to recruit subjects for this study. We wish to thank Robin Li, Coleen Hacker, Naiping Shen, Claire Marjoribanks, and Albert Yee for excellent technical assistance and overall contribution to this work. We thank Alberto Cevallos and Jesse Hsu, for helping to establish our Mexican collaboration, and Pascual Starink, for assistance with sample tracking. We also thank Kelly Frazer and two anonymous reviewers for helpful comments on the manuscript.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/ (for ss12673803–ss12674077)
Genetic Epidemiology Group Web Site, http://www.lshtm.ac.uk/eu/genetics/ (for ADMIXMAP software)
NCBI BLAST, http://www.ncbi.nlm.nih.gov/BLAST/ (for BLAST search engine)
NCBI Home Page, http://www.ncbi.nlm.nih.gov/
Pritchard Lab, http://pritch.bsd.uchicago.edu/ (for the *structure* program)
Q-VALUE Software, http://faculty.washington.edu/~jstorey/qvalue/
R Project for Statistical Computing, http://www.r-project.org/
RepeatMasker Web Server, http://ftp.genome.washington.edu/cgi-bin/RepeatMasker

References

Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. Am J Hum Genet 71:304–311
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 66:393–405
Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604
Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85:9119–9123
Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251:767–773
Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding in genetic associations in stratified populations. Am J Hum Genet 72:1492–1504
Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comp Graph Stat 5:299–314
Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) GM 3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520–526
Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144
Lander ES, Schork N (1994) Genetic dissection of complex traits. Science 265:2037–2048
Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723
McKeigue PM (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations by conditioning on parental admixture. Am J Hum Genet 63:241–51
McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet 64:171–186
Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228
Pritchard JK, Stephens M, Donnelly P (2000*a*) Inference of population structure using multilocus genotype data. Genetics 155:945–959
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000*b*) Association mapping in structured populations. Am J Hum Genet 67:170–181
Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16
Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–477
Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. Nat Rev Genet 3:862–871
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445
Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES 4th (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289.
Weir B (1996) Genetic data analysis II. Sinauer, Sunderland, MA