European Journal of Human Genetics (2004) 12, 1001–1006
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

ARTICLE

Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip® Human Mapping 10K array

Ke Hao[1], Cheng Li[1,2], Carsten Rosenow[3] and Wing H Wong*[,1,4]

[1]*Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA;* [2]*Department of Biostatistics, Dana Farber Cancer Institute, Boston, MA, USA;* [3]*Genomics Collaboration, Affymetrix, Santa Clara, CA, USA;* [4]*Department of Statistics, Harvard University, Cambridge, MA, USA*

Population-based association design is often compromised by false or nonreplicable findings, partially due to population stratification. Genomic control (GC) approaches were proposed to detect and adjust for this confounder. To date, the performance of this strategy has not been extensively evaluated on real data. More than 10 000 single-nucleotide polymorphisms (SNPs) were genotyped on subjects from four populations (including an Asian, an African-American and two Caucasian populations) using GeneChip® Mapping 10 K array. On these data, we tested the performance of two GC approaches in different scenarios including various numbers of GC markers and different degrees of population stratification. In the scenario of substantial population stratification, both GC approaches are sensitive using only 20–50 random SNPs, and the mixed subjects can be separated into homogeneous subgroups. In the scenario of moderate stratification, both GC approaches have poor sensitivities. However, the bias in association test can still be corrected even when no statistical significant population stratification is detected. We conducted extensive benchmark analyses on GC approaches using SNPs over the whole human genome. We found GC method can cluster subjects to homogeneous subgroups if there is a substantial difference in genetic background. The inflation factor, estimated by GC markers, can effectively adjust for the confounding effect of population stratification regardless of its extent. We also suggest that as low as 50 random SNPs with heterozygosity $>$40% should be sufficient as genomic controls.

European Journal of Human Genetics (2004) **12,** 1001–1006. doi:10.1038/sj.ejhg.5201273
Published online 15 September 2004

Keywords: population stratification; population-based study; association test; genomic control

Introduction

In theory, for equivalent sample size, association test is far more powerful than pedigree-based linkage studies in searching for genomic regions underlying human diseases.[1] The basic idea behind association test is that the disease alleles are more frequent in ascertained cases than in controls. Markers physically close to the disease loci will also be detected because of linkage disequilibrium (LD). However, the application of this approach is compromised by false or nonreplicable findings,[2] partially due to population stratification, which causes unlinked markers to show association with the phenotype.[3,4] Recent population admixture also bias association test, and an example is the spurious finding between immunoglobulin haplotype

*Correspondence: Dr WH Wong, Department of Biostatistics, Harvard University, Harvard School of Public Health, 655 Huntington Ave., Building II, Room 441, Boston, MA 02115, USA. Tel: +1 617 432 4912; Fax: +1 617 739 1781; E-mail: wwong@hsph.harvard.edu

Received 18 February 2004; revised 11 June 2004; accepted 22 July 2004

Gm[3,5,13,14] and NIDDM in the Gila River Indian Community.[5] The study was confounded by the subjects' degree of Caucasian genetic heritage.[6]

To overcome this serious danger, a correction strategy has been proposed.[6,7] It requires to genotype additional unlinked markers, often called 'genomic control (GC) markers', as the cost of detecting and correcting possible confounders. Under the assumption of no association between GC markers and phenotype and no population stratification, the χ^2 statistics of association test between the *i*th GC marker and case–control status, denoted as Y_i^2, follows a χ^2 distribution with one degree of freedom if using additive genetic model. And the sum of the χ^2 statistics of *n* GC markers, denoted as Y_n^2, follows a χ^2 distribution with *n* degree of freedom, where we can easily test whether the population stratification is present. Furthermore, we assume the test statistic is inflated by a factor λ, $Y_n^2/\lambda \sim \chi_n^2$. If we assume λ is constant for all loci, we can then use it to adjust the population stratification. One robust way to estimate the inflation factor is:

$$\lambda = Median\,(Y_i^2)/0.456,$$

where 0.456 is simply the median of χ^2 distribution with one degree of freedom.[8] We denote this method as the combined χ^2 approach in this paper. An alternative method, proposed by Pritchard *et al*,[9] tackles this problem in two steps. Firstly, the GC markers are used to separate the study subjects into genetically homogeneous subgroups, and second, the association tests will be conducted within each subgroup.

To date, the performance of the GC approaches has not been examined extensively in real genotype data. Previous researches were based on simulated data or small number of GC markers. The Affymetrix Mapping 10 K array has recently become available, offering the ability to genotype more than 10 000 single-nucleotide polymorphisms (SNPs) across the human genome in a timely manner.[10] Using this technology, we evaluated the current genomic control approaches in testing and controlling for population stratification.

Methods

Study subjects

Four groups of subjects were used in the current study, (1) 20 Asians, (2) 42 African-Americans, (3) 42 Caucasian collected by Coriell Institute, and (4) 54 Caucasian subjects collected from Utah, USA. The DNA samples of group (1–3) were purchased from Coriell Institute, and the group (4) samples were collected by Centre d'Etude du Polymorphisme Humain (CEPH) research laboratory. All subjects were unrelated individuals and remained anonymous to the authors.

Genotyping

A measure of 250 ng genomic DNA of each subject was digested with *Xba*I at 37°C for 4 h. The DNA fragments undergo ligation to a universal adaptor and then PCR-amplification with a common primer. The amplicon was cleaved by partial *Dnase*I digest to shorter fragments, and labeled with biotinylated ddATP using terminal deoxytransferase. The labeled DNA was injected into the microarray cartridge and incubated overnight. The hybridized microarray was washed and stained following a three-step protocol, and was scanned under the manufacturer's directions (Affymetrix). Finally, the genotype was determined using an automated scoring software (Affymetrix). The detailed genotyping procedure used has been previously described elsewhere.[10] This data set has been made available to the public at http://www.affymetrix.com/support/developer/resource_center/index.affx?terms = no.

Statistical analysis

Only autosomal markers were used in the analysis. We firstly compared the allele frequencies and heterozygosities of the genotyped SNPs among populations. Second, we evaluated the performance of genomic control method in detecting population stratification through an iterative procedure. We pooled genotypes from different groups together, that is, Asian and African Americans, and attempted to detect this mixture using GC approach. In each iteration loop, we randomly selected $n = 20$ or 50 SNPs from the data set, calculated Y_n^2 in the combined χ^2 approach, and conducted test for population stratification. Overall, 10 000 iterations were carried out, and we summarized the power as P ($P < 0.05$). Furthermore, we assessed the degree of bias it would cause in association tests if we ignored the underlying population stratification. The Armitage's trend test for additive model was used.[8] We randomly assigned a fraction (0, 25, 50, 75 and 100%) of each ethnic group to be disease affected, pooled two groups together, and tested for disease–SNP association. This simulation procedure was repeated 10 000 times, and we recorded the rejection rate at 0.05 level. Upon observed substantial population stratification, we also estimated the inflation factor (λ), and calculated the rejection rate again with controlling for population stratification.

We also applied the Pritchard's approach, which is a model-based clustering method using unlinked SNPs to infer population structures, and assign individuals to clusters.[9] The method is implemented in a software, *STRUCTURE* (version 2), which was downloaded from http://pritch.bsd.uchicago.edu. We evaluated this method by pooling two ethnic groups together, and run STRUCTURE to detect the population structure using 50 or 500 GC SNPs.

Results

A total of 158 unrelated individuals from four ethnic groups were genotyped on 10 043 SNP markers by the array, with an overall call rate of 96.4%. These SNPs are fairly polymorphic in our study samples, and the average heterozygosity (>40%) and allele frequency (>20%) of the SNPs were similar across all four ethnic groups (Table 1). Using the combined χ^2 approach, we found 10–20 SNPs were sufficient to detect population stratification in the scenarios of Asian-Caucasian, Asian-African American and Caucasian-African American mixture at the nominal 0.05 level (Table 2). The power in rejecting the null (no population stratification) was over 95% in these cases by only using 10 genomic control SNPs. However, when mixing the Caucasian subjects collected by Coriell Inc. and those collected by CEPH, we have limited power to detect the stratification even using 50 random SNPs (Table 2). One possibility stands as there was no significant population stratification between these two groups of Caucasian subjects, so that we can conduct association test without adjustment. However, we observed substantial bias in the test if we mix any two ethnic groups together (Table 3). In the cases of mixing the two groups of Caucasian subjects, the rejection rate could be more than 20% under the null hypothesis (Table 3). It should be noted, that in the 0.5/0.5 situation of Table 3, we simulated case–control studies matched on ethnicity. When sample size is small-to-moderate, using asymptotic χ^2 distribution tends to yield overestimated *P*-value and result into conservative test.[11] Only when sample size becomes large, the asymptotic *P*-value is accurate.[11] As a consequence, in the 0.5/0.5 column of Table 3, the rejection rates were slightly less than α level except when mixing the two Caucasian groups where population stratification was less severe. Upon observing strong bias in marker–disease association testing if ignoring the population stratification, we used the estimated inflation factor (λ) to adjust the association tests, and obtained correct rejection rate (Table 4). In addition, we simulated situations of mixing two ethnic groups (eg Asian and African-American), where one group was matched in cases and controls but the other group was mismatched. In this case, we also observed elevated false-

Table 1 Mean heterozygosity and allele frequency of the SNPs among study subjects

Groups	*Caucasian (n=42)*	*Utah (n=54)*	*Asian* (n=20)	*African-American (n=42)*
Heterozygosity (%)	45.9	45.8	41.3	46.8
Allele frequency (%)	25.3	25.0	22.8	25.2

Table 2 Power of testing for population stratification at 0.05 level*

	10 random SNPs		*20 random SNPs*		*50 random SNPs*	
Ethnic groups	*Power*	*M(p)*	*Power*	*M(p)*	*Power*	*M(p)*
Asian *vs* Caucasian	97.2%	3.5×10^{-5}	100%	3.3×10^{-5}	100%	$<10^{-15}$
Utah *vs* Caucasian	9.0%	0.424	23.2%	0.192	38.8%	0.091
African-American *vs* Caucasian	99.2%	2.9×10^{-11}	100%	$<10^{-15}$	100%	$<10^{-15}$
Asian *vs* Utah	97.4%	2.8×10^{-8}	100%	$<10^{-15}$	100%	$<10^{-15}$
African-American *vs* Asian	99.4%	3.6×10^{-9}	99.9%	$<10^{-15}$	100%	$<10^{-15}$
African-American *vs* Utah	99.8%	6.5×10^{-13}	100%	$<10^{-15}$	100%	$<10^{-15}$

*Power is estimated on 10 000 iterations; $M(p)$, median P-value.

Table 3 Rejection rate of association test under the null hypothesis at 0.05 level[a]

Case/control ratio	*1/0 (%)*	*0.75/0.25 (%)*	*0.5/0.5 (%)*	*0.25/0.75 (%)*	*0/1 (%)*
Asian *vs* Caucasian	55.4	28.1	4.1	28.0	56.8
Utah *vs* Caucasian	23.1	9.4	5.1	9.1	22.0
African American *vs* Caucasian	62.2	37.0	4.4	37.2	62.3
Asian *vs* Utah	57.5	28.9	4.2	27.2	57.6
African American *vs* Asian	60.9	32.5	4.3	32.5	60.2
African American *vs* Utah	66.2	40.0	4.3	40.7	65.4

[a] Estimation was based on 10 000 iterations. Caucasian, Caucasian samples collected by Coriell Institute. Utah, Caucasian samples collected by CEPH lab.

Table 4 Controlling for population stratification with 50 unlinked markers*

Case/control ratio	*1/0*	*0.75/0.25*	*0.6/0.4*	*0.5/0.5*	*0.4/0.6*	*0.25/0.75*	*0/1*
(a) Adjusted rejection rate in association test under the null hypothesis at 0.05 level							
Asian *vs* Caucasian	4.4%	4.3%	4.2%	3.5%	4.0%	4.3%	4.1%
Utah *vs* Caucasian	4.4%	4.1%	4.0%	3.8%	4.0%	4.0%	4.4%
African-American *vs* Caucasian	3.5%	3.9%	4.4%	3.6%	4.2%	4.0%	3.5%
Asian *vs* Utah	4.6%	4.6%	4.2%	3.6%	4.0%	4.5%	4.7%
African-American *vs* Asian	3.4%	4.0%	3.9%	3.4%	3.9%	3.1%	3.4%
African-American *vs* Utah	3.3%	3.9%	4.4%	3.5%	4.2%	4.0%	3.1%
(b) Mean (variance) of the inflation factor, λ							
Asian *vs* Caucasian	5.67 (3.63)	2.20 (0.46)	1.26 (0.10)	1.14 (0.05)	1.24 (0.30)	2.07(0.40)	5.84 (3.93)
Utah *vs* Caucasian	1.22 (0.08)	1.17 (0.06)	1.16 (0.06)	1.18 (0.06)	1.16 (0.06)	1.16(0.05)	1.21 (0.08)
African-American *vs* Caucasian	8.62 (7.12)	2.92 (0.82)	1.38 (0.14)	1.12 (0.04)	1.39 (0.17)	2.91 (0.77)	8.62 (7.12)
Asian *vs* Utah	6.49 (4.83)	2.23 (0.50)	1.27 (0.10)	1.12 (0.04)	1.24 (0.09)	2.14 (0.41)	6.52 (4.73)
African-American *vs* Asian	7.51 (5.39)	2.42 (0.56)	1.29 (0.11)	1.11 (0.04)	1.34 (0.12)	2.95 (0.87)	7.72 (6.83)
African-American *vs* Utah	10.1 (9.93)	3.24 (1.10)	1.36 (0.12)	1.12 (0.04)	1.44 (0.18)	3.25 (1.09)	10.5 (12.1)

Estimation was based on 10 000 iterations, the rejection rates were summarized in panel (a). In panel (b), we presented the mean value and variance of the 10 000 λs obtained in the 10 000 times simulation. Caucasian, Caucasian samples collected by Coriell Institute. Utah, Caucasian samples collected by CEPH lab.

positive rate caused by population stratification, which could be appropriately adjusted for by using χ^2 genomic control methods.

In the cases of Asian-Caucasian, Asian-African American and Caucasian-African American pooling, the STRUCTURE software can easily separate the two groups using 50 random SNPs. In contrast, when we combine the two groups of Caucasian samples, the software failed to detect the population stratification even using 500 SNPs. Because the computation time of this method is fairly long, we carried out only 20 runs using different sets of 500 random SNPs, but in none of the 20 occasions the subpopulations were detected.

Discussion

In this paper, we evaluated two different strategies in detecting and controlling for population stratification using real data. We surveyed the cases of (1) pooling genetically distant populations, such as Asians and Caucasians, and (2) pooling genetically similar populations, such as two groups of Caucasian samples. In case (1), both strategies were able to detect the stratification with a small number of GC SNPs, however, in case (2), the sensitivity is low in both strategies even with hundreds of SNPs. The inflation factor (λ) in the combined χ^2 approach can correctly adjust the confounding effect even when the population stratification was statistically nonsignificant. In contrast, no adjustment can be attempted in Pritchard's approach when subpopulation structure is not detected.

In the past decade, we have witnessed the rise of family-based designs as an alternative to population-based study.[8] The motivation is that family-based designs are protected from population stratification by its nature.[12] This protection comes with a cost: (1) family-based samples are more difficult to collect, and (2) conditioning on the same number of genotypes, family-based tests are less powerful than their population-based counterparts.[13] Furthermore, the genomic control approaches made population-based study at least comparable to family-based tests.

We can generally consider two scenarios. (1) In the situation of substantial population stratification, Pritchard's method appears to be the most attractive. It can cluster subjects into homogeneous groups, and tests can be conducted within these groups. We should note that, in this case, the gene–disease association could be quite different among ethnic groups in terms of both magnitude and direction due to the distinct genetic backgrounds. Family-based approach can only detect the average genetic effect, which could miss the association if its directions are opposite among subpopulations. (2) In the situation of subtle population stratification, such as two groups of Caucasian subjects collected from different geological regions, Pritchard's method showed limited sensitivity in detecting the stratification. Fortunately, the combined χ^2 approach can still appropriately adjust for the confounding effect using estimated λ. Our results suggest that the adjustment is fairly accurate in various degrees of population stratification. Simulation studies also showed, in this case, population-based study with GC adjustment is statistically more powerful than family-based tests.[8]

The degree of population stratification varies among genetic markers. Some markers carry similar allele frequencies across populations, and in contrast, some markers are ethnic specific. When the two underlying populations are not separable (ie, the scenario of mixing two Caucasian samples), we have to estimate the λ on a number of random

SNPs and apply it as a constant. By these means, we controlled the overall false-positive rate on all SNPs. However, for each individual SNP, its degree of population stratification could be under- or overestimated. The advantage of the χ^2 approach is that it can adjust for population stratification even when the underlying populations are not separable. Its drawback is, when we are able to separate the underlying populations, this method is not the most efficient way to adjust for the stratification, on the other hand, the STRUCTURE strategy is more reasonable in this scenario. The disadvantage of the STRUCTURE strategy is, when the underlying populations are not separable, it simply cannot provide any adjustment.

It should be noted that, in the situation of subtle population stratification, both the combined χ^2 approach and the Pritchard's method showed limited power in detecting it. However, we may still suffer from severe biases if the association test was performed without adjustment. In this study, we found the inflation factor can solve this potential danger, even when no significant subpopulations are detected. Hence, adjusting the test using λ is recommended in a population-based association study if GC data are available. In addition, we set $\lambda \geq 1$, which partially caused that, Table 4a, the corrected rejection rate under the null hypothesis is slightly less than 5%.

'How many GC markers should we use?' is always an intriguing question, and different suggestions have been made.[7–9] Actually, the answer to this question highly depends on the population structure, sample size and heterozygosity of genomic control makers, and these parameters varies greatly across studies. Thus, no simply cutoff can be suggested. In this study, we investigated the impact of allele frequencies of genomic control SNPs to the power in detecting the stratification. We found frequent SNPs provide much larger power than infrequent SNPs. In our data, SNPs with minor allele frequency less than 15% offer nearly no power. Here, 50 SNPs with an average heterozygosity around 40% provided great power to detect population stratification and to make appropriate adjustment. As demonstrated in Table 4b, the variance of λ is fairly small, which means no matter which 50 SNPs on the genome we choose as GC marker they will lead to similar estimations of λ. With the rapid progress of biotechnology, the cost of SNP genotyping has been reduced greatly. To genotype a set of 50 SNPs in a population-based study is no longer a major financial or technological hurdle (in comparison to collecting family samples); moreover, GC methods will provide valuable and often necessary adjustments. GC approach can also be applied to family-based tests. When the direction of association is opposite among subpopulations, family-based tests may lead to a mistaken null finding. Since GC markers can separate sample to genetically homogeneous subgroups, conducting family-based test within these groups is arguably more appropriate and powerful.

In another setting, where we would like to examine whether a group of individuals belongs to a certain population (eg Asian), using known ethnic-specific markers would be more powerful and efficient than random markers. To date, numbers of these ethnic-specific markers have been characterized on many populations.[14]

After our initial submission to the *European Journal of Human Genetics*, two papers on this topic were published in the *Nature Genetics*.[11,14] Here, we take the opportunity of manuscript revision to compare the designs and results of these studies. Freedman *et al*[14] utilized multiple populations of moderate-to-large sample size, but only typed a few dozen markers on each sample. They observed similar results as ours, that subtle population stratification is not detectable with adequate power by using χ^2 methods. However, in this situation, the subtle population stratification still increases the likelihood of false positives. Unfortunately, Freedman *et al* looked at neither the usage of λ in adjusting association test or the STRUCTURE strategy in detail. Similar to our study, Marchini *et al*[11] typed large number of SNPs on small-to-moderate sample sizes. Using a Bayesian method,[11,15] they found substantial stratification among Asian, White and Black subjects, and much smaller difference between Chinese and Japanese. Furthermore, Marchini *et al*[11] simulated large cohorts and investigated the impact of sample size on association tests in the context of population stratification. Their results agree well with our findings.

Acknowledgements

We would like to thank Dr Rui Mei for providing the data set and running the arrays. We thank Dr Xin Xu and Dr Tianhua Niu for carefully reading the manuscript and providing insightful comments. This work is partially supported by NIH Grant 1R01HG02341.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Weiss ST, Silverman EK, Palmer LJ: Case–control association studies in pharmacogenetics. *Pharmacogenom J* 2001; **1**: 157–158.
3 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 1994; **265**: 2037–2048.
4 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57**: 455–464.
5 Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988; **43**: 520–526.
6 Thomas DC, Witte JS: Point: population stratification: a problem for case–control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 2002; **11**: 505–512.
7 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
8 Bacanu SA, Devlin B, Roeder K: The power of genomic control. *Am J Hum Genet* 2000; **66**: 1933–1944.

9 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.

10 Kennedy GC, Matsuzaki H, Dong S *et al*: Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003; **21**: 1233–1237.

11 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.

12 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.

13 Morton NE, Collins A: Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci USA* 1998; **95**: 11389–11393.

14 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.

15 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.