Published in final edited form as:

Sci China C Life Sci. 2009 October ; 52(10): 972–976. doi:10.1007/s11427-009-0128-y.

A PCA-based method for ancestral informative markers selection in structured populations

ZHANG Feng^{1,3}, ZHANG Lei^{1,3}, and DENG Hong-Wen^{1,2,3,†}

¹Key Laboratory of Biomedical Information Engineering of Ministry of Education, Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China;

²Institute of Bioscience and Biotechnology, School of Science, Beijing Jiaotong University, Beijing 100044, China;

³Departments of Orthopedic Surgery and Basic Medical Science, School of Medicine, University of Missouri-Kansas City, Kansas City, MO 64108, USA

Abstract

Identification of population structure can help trace population histories and identify disease genes. Structured association (SA) is a commonly used approach for population structure identification and association mapping. A major issue with SA is that its performance greatly depends on the informativeness and the numbers of ancestral informative markers (AIMs). Present major AIM selection methods mostly require prior individual ancestry information, which is usually not available or uncertain in practice. To address this potential weakness, we herein develop a novel approach for AIM selection based on principle component analysis (PCA), which does not require prior ancestry information of study subjects. Our simulation and real genetic data analysis results suggest that, with equivalent AIMs, PCA-based selected AIMs can significantly increase the accuracy of inferred individual ancestries compared with traditionally randomly selected AIMs. Our method can easily be applied to whole genome data to select a set of highly informative AIMs in population structure, which can then be used to identify potential population structure and correct possible statistical biases caused by population stratification.

Keywords

population structure; principle component analysis; ancestral informative markers

Population structure is a common feature in human populations[1,2]. Identification of population structure can help trace population histories in population genetics[3]. Furthermore, identifying population structure is effective for correcting population stratification in association studies of human diseases[4–6].

Structured association (SA) is a commonly used approach for population structure identification and association mapping[5,7–9]. SA uses a set of genetic markers, known as ancestral informative markers (AIMs), to infer individual ancestries, which can then be used to test disease-gene association while correcting population stratification[4,5]. A major issue with SA is that its performance greatly depends on the informativeness and the numbers of AIMs[10,11], which may limit its robustness and efficiency in practical

applications. Due to the extensive computational demand by SA[12,13], it is generally difficult to improve the performance of SA by simply increasing the numbers of AIMs, especially in the large-scale studies with thousands of subjects and hundreds of thousands of genetic markers. Therefore, selecting a set of highly informative AIMs for SA is an alternative and complementary solution.

Several AIM selecting methods are currently available[1,10]. For instance, AIMs can be selected to maximize absolute allele frequency differences among different ancestral populations, or Wright's F_{ST} [1,10]. However, these methods require prior knowledge about individuals' membership or ancestries to known populations, which are usually not available or uncertain in practice. Therefore, it is difficult to apply these methods to structured populations without prior ancestry information of study subjects.

Principle component analysis (PCA) is a classical dimensionality reduction technique and has been used in genetic studies[6,9,12,14–16]. In this paper, we introduce a novel PCA-based approach for AIM selection, which does not require prior ancestry information of study subjects. We simulated a set of stratified populations based on real haplotype data from the HapMap ENCODE project, and evaluated the accuracy of inferred individual ancestries using PCA-based and randomly selected AIMs.

1 Materials and methods

1.1 PCA-based algorithm for AIM selection

Our method encompasses two steps: first, principle component analysis is applied to genotypic data to infer continuous variation axes of population structure, using individuals as variables and markers as samples[12]. Supposed genotypic data of M individuals are collected. For a given variation axis i , the ancestry of individual j ($j=1, 2, \dots, M$) can be denoted by the j th coordinate of the eigenvector of variation axis i [12]. Second, we construct an information measure of population structure S_{il} , defined by

$$S_{il} = \text{Var}(a_i \times g_{.l}),$$

where a_i is the $M \times 1$ eigenvector of variation axis i , and $g_{.l}$ is a $1 \times M$ vector of genotype at SNP l . Because significant variation axes denote differentiation directions of population structure[12], SNPs with larger S_{il} are reasonable to be more informative about population structure than SNPs with smaller S_{il} . For the i th variation axis, we can select n_i SNPs, which attain larger S_{il} than remaining SNPs and have not yet been selected by previous $i-1$ variation axes ($i=2, 3, \dots$). n_i is defined by

$$n_i = \left\lceil N \times \frac{v_i}{V} \right\rceil,$$

where N denotes the total number of AIMs needing to select, v_i denotes the variance explained by the i th variation axis and V denotes the total variance of analyzed data in PCA. The above procedure will continue until sufficient AIMs are selected.

1.2 Simulations

A simulation study was conducted to evaluate the relative performance of PCA-based selected AIMs and traditionally randomly selected AIMs for population structure identification in structured populations without prior ancestry information of study subjects. Simulating algorithm has been detailed in our earlier study[17]. Briefly, phased haplotype

data of Caucasians with northern and western European ancestry (CEPH) and Yoruba from Ibadan (YRI) of Africa were downloaded from HapMap ENCODE website (http://www.HapMap.org/downloads/phasing/2005-03_phaseI/ENCODE). Within each ENCODE region, we selected the set of highly informative marker loci, which were genotyped in both CEPH and YRI and were polymorphic in at least one population or monomorphic, but had different alleles in the two populations. There were 12867 highly informative marker loci selected from 10 ENCODE regions. We converted the genetic map distances reported by the HapMap ENCODE project to recombination fractions between adjacent informative marker loci using the Kosambi map function[18]. Based on the phased CEPH and YRI haplotype data and derived recombination fractions, CEPH and YRI subpopulations were first separately simulated, and then mixed together to generate a structured populations with 300 CEPH individuals and 100 YRI individuals.

To model low (Model 1), modest (Model 2) and high (Model 3) population stratification levels, 500 of 12867 markers were first selected with preset distribution of allele frequency differences between the simulated CEPH and YRI subpopulations (Table 1). PCA-based selected and traditionally randomly selected AIMs were then selected from the 500 markers. The numbers of AIMs were varied to assess the relative performance of PCA-based selected AIMs and traditional randomly selected AIMs.

STRUCTURE 2.1 was applied to infer individual ancestries using PCA-based selected AIMs and randomly selected AIMs[4,5]. For simplicity, the presumed number of subpopulations of the simulated structured populations used by STRUCTURE 2.1 was pre-assigned 2 in all models. STRUCTURE 2.1 was running under default parameters recommended by the program developers. True individual ancestries were identified from the simulations. The true ancestral proportions of YRI individuals were assigned 1.0, and the true ancestral proportions of CEPH individuals were assigned 0.0. The true ancestral proportions were regressed on the ancestral proportions estimated by STRUCTURE 2.1[10]. Because of the extensive computational cost required by SA[12,13], 200 simulations were conducted for each model. Average regression coefficients in the 200 simulations were calculated and used to evaluate the accuracy of inferred individual ancestries[10].

1.3 Application to real genetic data

To evaluate the performance of our method, we further analyzed real genetic data, which came from USA Framingham Heart Study (<http://www.framingham-heartstudy.org/gen/index.html>) and our experimental data. Briefly, the genotype data for 500 unrelated USA whites at chromosome 1 were first retrieved from the Framingham Heart Study database, which was publicly available (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v4.p2). We then randomly selected 500 unrelated Han Chinese subjects from our genotype database. The information of our Han Chinese samples was detailed in our previous paper[19]. To improve the precision of our study, strict quality control for genotyping data was employed to both USA whites and Han Chinese genotype data, discarding the SNPs with minor allele frequency (MAF) <1%, and deviating from Hardy-Weinberg equilibrium (HWE) ($P < 0.001$). 29956 SNPs at chromosome 1, which were genotyped in both USA whites and Han Chinese, were used in the following analysis. The USA whites and Han Chinese were mixed together to generate a structured population with 1000 unrelated individuals. The similar to the above simulation study, PCA-based algorithm for AIMs selection was applied to the real structured population, and compared with traditionally randomly selected AIMs under various numbers of AIMs.

2 Results

Simulation results are summarized in Figure 1. As expected, the accuracy of inferred individual ancestries increased with increasing numbers of AIMs. PCA-based selected AIMs generally performed better than traditionally randomly selected AIMs under various scenarios (Figure 1). Under the low population stratification level (Model 1), with equivalent AIMs, the PCA-based selected AIMs performed significantly better than the randomly selected AIMs. Under Models 2 and 3 with modest and high population stratification levels, the performance of the randomly selected AIMs became close to that of the PCA-based selected AIMs, only when 100 AIMs were used.

Figure 2 presents the analysis results of real genetic data. In the real structured population, PCA-based selected AIMs showed higher accuracy of inferred individual ancestries compared with traditionally randomly selected AIMs under various numbers of AIMs, which is consistent with our simulation results.

3 Discussion

It has been demonstrated that AIM selection can significantly affect the performance of SA[10,11]. Because the ancestry information of study subjects is usually not available or uncertain in practice, it is difficult to select AIMs based on prior individual ancestry information. To address this issue, we developed a novel PCA-based AIMs selecting approach, which does not require prior individual ancestry information. Our simulation suggests that PCA-based selected AIMs can increase the accuracy of inferred individual ancestries compared with traditional randomly selected AIMs, especially under subtle population stratification that is common in human populations[20]. Additional real genetic data analysis results support the improved performance of our novel method in individual ancestry inference. The increased accuracy of inferred individual ancestries can further improve the performance of SA correcting population stratification in association studies of human diseases[4,5].

Currently, with the rapid development of high-throughput genotyping technologies, large-scale association studies are popular in gene mapping of human diseases, such as osteoporosis, heart diseases and diabetes[21–24]. Nonetheless, an outstanding issue complicating PBAS is population structure, which can cause spurious association results and limit the robustness and efficiencies of PBAS. Our method can be easily applied to whole genome data to select a set of highly informative AIMs in population structure, which can then be used to identify potential population structure and correct possible statistical biases caused by population stratification.

Four aspects from our study may be noted. First, we used STRUCTURE to evaluate the performance of our novel method for population structure identification. Our method can also be directly applied to other population structure identification and association mapping approaches, for example admixture mapping[25–28]. Second, the real CEPH and YRI haplotype data from the Hap-Map ENCODE project were used to simulate genotype data for each subject. The simulated data sets are hence close to the realistic scenario, which ensures the robustness of our simulation results. Third, we used the variance proportions explained by each population structure variation axis in PCA to determine the numbers of AIMs selected for each variation axis. This approach tends to select more AIMs for more significant variation axes of population structure, which may increase the accuracy of inferred individual ancestries. Fourth, to ensure the effectiveness and robustness of our simulation approach for population structure, we selected 12867 marker loci from the 10 ENCODE regions to conduct simulation. The 12867 marker loci are highly informative

about population structure, which may compromise the performance improvement of PCA-based selected AIMs compared with traditional randomly selected AIMs.

We introduce a novel algorithm for AIMs selection based on principle component analysis (PCA), which does not require prior ancestry information of study subjects. Our methods can be easily applied to whole genome data to identify potential population structure and correct for possible statistical bias caused by population structure.

Acknowledgments

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI.

Supported by Xi'an Jiaotong University, NIH (Grant Nos. R01 AR050496, R21 AG 027110, R01 AG026564 and P50 AR055081), Fok Ying Tung Education Foundation and Framingham Heart Study and the Framingham SHARe Project

References

1. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;36:388–393. [PubMed: 15052270]
2. Guthery SL, Salisbury BA, Pungliya MS, et al. The structure of common genetic variation in United States populations. *Am J Hum Genet* 2007;81:1221–1231. [PubMed: 17999361]
3. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 2003;33(Suppl):266–275. [PubMed: 12610536]
4. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155:945–959. [PubMed: 10835412]
5. Pritchard JK, Stephens M, Rosenberg NA, et al. Association mapping in structured populations. *Am J Hum Genet* 2000;67:170–181. [PubMed: 10827107]
6. Zhu X, Li S, Cooper RS, et al. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* 2008;82:352–365. [PubMed: 18252216]
7. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. *Science* 2002;298:2381–2385. [PubMed: 12493913]
8. Lao O, Van Duijn K, Kersbergen P, et al. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet* 2006;78:680–690. [PubMed: 16532397]
9. Chen HS, Zhu X, Zhao H, et al. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet* 2003;67:250–264. [PubMed: 12914577]
10. Tsai HJ, Choudhry S, Naqvi M, et al. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Hum Genet* 2005;118:424–433. [PubMed: 16208514]
11. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001;60:227–237. [PubMed: 11855957]
12. Price, a L.; Patterson, NJ.; Plenge, RM., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909. [PubMed: 16862161]
13. Liu N, Zhao H. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics* 2006;2:353–364. [PubMed: 16848973]
14. Zhu X, Zhang S, Zhao H, et al. Association mapping, using a mixture model for complex traits. *Genet Epidemiol* 2002;23:181–196. [PubMed: 12214310]

15. Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol* 2003;24:44–56. [PubMed: 12508255]
16. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science* 1978;201:786–792. [PubMed: 356262]
17. Zhang F, Liu J, Chen J, et al. HAPSIMU: A genetic simulation platform for population-based association studies. *BMC Bioinformatics* 2008;9:331. [PubMed: 18681963]
18. Kosambi DD. The estimation of map distances from recombination values. *Ann Eugen* 1944;12:172–175.
19. Yang TL, Chen XD, Guo Y, et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet* 2008;83:663–674. [PubMed: 18992858]
20. Marchini J, Cardon LR, Phillips MS, et al. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;36:512–517. [PubMed: 15052271]
21. Matarin M, Brown WM, Singleton A, et al. Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. *Stroke* 2008;39:1586–1589. [PubMed: 18340101]
22. Sklar P, Smoller JW, Fan J, et al. Whole-genome association study of bipolar disorder. *Mol Psychiatry* 2008;13:558–569. [PubMed: 18317468]
23. Salonen JT, Uimari P, Aalto JM, et al. Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. *Am J Hum Genet* 2007;81:338–345. [PubMed: 17668382]
24. Liu YZ, Wilson SG, Wang L, et al. Identification of PLCL1 gene for hip bone size variation in females in a genome-wide association study. *PLoS ONE* 2008;3:e3160. [PubMed: 18776929]
25. Mckeigue PM. Prospects for admixture mapping of complex traits. *Am J Hum Genet* 2005;76:1–7. [PubMed: 15540159]
26. Mckeigue PM, Carpenter JR, Parra EJ, et al. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* 2000;64:171–186. [PubMed: 11246470]
27. Zhu X, Tang H, Risch N. Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet* 2008;60:547–569. [PubMed: 18358332]
28. Zhu X, Zhang S, Tang H, et al. A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* 2006;120:431–445. [PubMed: 16896924]

Figure 1. Regression analysis results of the ancestral proportions in simulated structured populations using PCA-based selected AIMs and randomly selected AIMs.

Figure 2. Regression analysis results of the ancestral proportions in real structured population using PCA-based selected AIMs and randomly selected AIMs.

Table 1

Distribution of allele frequency differences (AFD) between CEPH and YRI subpopulations in simulated structured populations

	$ AFD \leq 0.02$	$0.02 < AFD \leq 0.10$	$0.10 < AFD \leq 0.30$	$0.30 < AFD \leq 0.50$
Model 1	90%	4%	4%	2%
Model 2	70%	12%	12%	6%
Model 3	50%	20%	20%	10%