Research article

Characterization of a likelihood based method and effects of markers informativeness in evaluation of admixture and population group assignment

Bao-Zhu Yang[1,2], Hongyu Zhao[3], Henry R Kranzler[4] and Joel Gelernter*[1,2]

Address: [1]Yale University School of Medicine, Department of Psychiatry, New Haven, CT, USA, [2]VA CT Healthcare Center, West Haven, CT, USA, [3]Yale University School of Medicine, Departments of Epidemiology and Public Health, and Genetics, New Haven, CT, USA and [4]University of Connecticut Health Center, Farmington, CT, USA

Email: Bao-Zhu Yang - bao-zhu.yang@yale.edu; Hongyu Zhao - hongyu.zhao@yale.edu; Henry R Kranzler - kranzler@psychiatry.uchc.edu; Joel Gelernter* - joel.gelernter@yale.edu

* Corresponding author

Abstract

Background: Detection and evaluation of population stratification are crucial issues in the conduct of genetic association studies. Statistical approaches useful for understanding these issues have been proposed; these methods rely on information gained from genotyping sets of markers that reflect population ancestry. Before using these methods, a set of markers informative for differentiating population genetic substructure (PGS) is necessary. We have previously evaluated the performance of a Bayesian clustering method implemented in the software STRUCTURE in detecting PGS with a particular informative marker set. In this study, we implemented a likelihood based method (LBM) in evaluating the informativeness of the same selected marker panel, with respect to assessing potential for stratification in samples of European Americans (EAs) and African Americans (AAs), that are known to be admixed. LBM calculates the probability of a set of genotypes based on observations in a reference population with known specific allele frequencies for each marker, assuming Hardy Weinberg equilibrium (HWE) for each marker and linkage equilibrium among markers.

Results: In EAs, the assignment accuracy by LBM exceeded 99% using the most efficient marker FY, and reached perfect assignment accuracy using the 10 most efficient markers excluding FY. In AAs, the assignment accuracy reached 96.4% using FY, and >95% when using at least the 9 most efficient markers. The comparison of the observed and reference allele frequencies (which were derived from previous publications and public databases) shows that allele frequencies observed in EAs matched the reference group more accurately than allele frequencies observed in AAs. As a result, the LBM performed better in EAs than AAs, as might be expected given the dependence of LBMs on prior knowledge of allele frequencies. Performance was not dependent on sample size.

Conclusion: The performance of the LBM depends on the efficiency and number of markers, and depends greatly on how representative the available reference allele frequencies are for those of the population being assigned. This method is of value when the parental population is known and relevant allele frequencies are available.
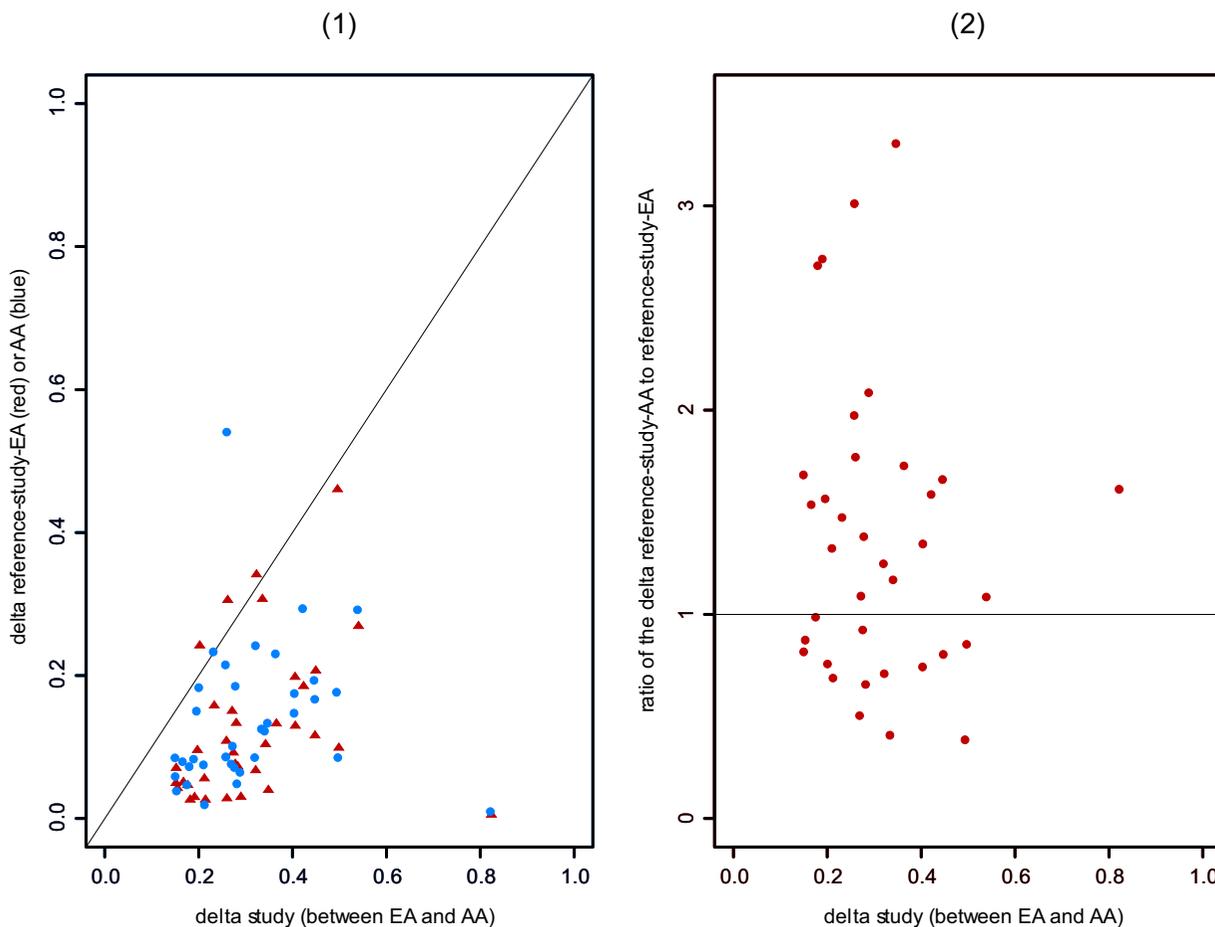
Figure 1
Marker efficiency in terms of the metric δ. (1) Comparison of delta between for AAs and EAs as observed in our sample, and as reported in the prior literature: $\delta_{study-AA-EA}$ versus $\delta_{reference-study-EA}$ (red triangle) or $\delta_{reference-study-AA}$ (blue dot). (2) Ratio for deltas for each marker (AA/EA) as observed in our sample compared to the prior literature: $\delta_{study-AA-EA}$ versus the ratio of $\delta_{reference-study-AA}$ to $\delta_{reference-study-EA}$.

Background

Population stratification is a crucial issue in conducting genetic association studies, in particular, for case-control study designs, such that if it is not accounted for study results could be invalid – either false positive or false negative [1]. Methods to address the issue have been proposed [2-17]. Before using these methods, an informative set of markers is necessary; this is known as a set of ancestry informative markers (AIMs). In this study, we implemented a likelihood based method (LBM), as an alternative to popular Bayesian methods such as that implemented in STRUCTURE [3,13], and used it to evaluate the informativeness of a selected marker panel and to assess potential for stratification in a sample of European

Americans (EAs) and African Americans (AAs) that are known to be admixed.

Likelihood-based methods (LBMs) provide a framework for assignment of individuals to specific populations based on observed allele frequencies in AIMs. LBMs for the classification of individuals into subgroups can be implemented by calculating the probability of a marker genotype profile (i.e., a set of genotypes) based on observations in a reference population with known specific allele frequencies for each marker ("training frequencies"), assuming Hardy Weinberg equilibrium (HWE) for each marker and linkage equilibrium among markers [18]. The LBM method is also called an "assignment test"
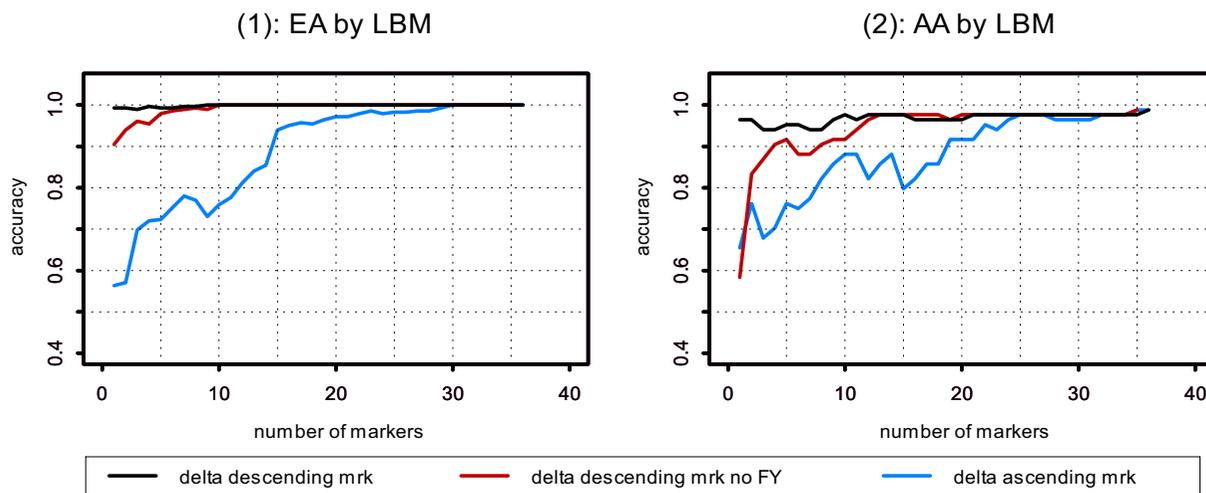
Figure 2
Assignment accuracy by LBM. Assignment accuracy by LBM. The markers are adding one by one either by δ descending or ascending. Assignment accuracy without FY, the most efficient marker in the panel studied, was also evaluated.

and is widely applied in molecular ecology and animal forensics for identifying population genetic substructures for animals or plants [18-24]. Research on the assignment test or LBM has not yet focused on the performance of the test or of specific markers in differentiating the PGS in human subjects. In theory, LBM may be better for probabilistic classification of individuals to subpopulations, if certain conditions are met. The most important of these conditions is availability of an accurate set of training frequencies. Obviously, this method may be applicable only if the populations from which the sample to be classified are already known or can be determined. This condition can be met in most situations; for example, the AA population is well known to have principally African and European American ancestry.

In the present study, we compared the performance of LBMs to that of the popular Bayesian approach used by the software program STRUCTURE. We predicted that, if the conditions for successful LBM application are met, LBMs would be more efficient that Bayesian methods for population group assignment, because they make use of more information (i.e., known ancestral population allele frequencies, which are provided *a priori* rather than inferred from the data presented to the program).

Results
We calculated the measure of marker efficiency by the metric δ for each marker. (Note that δ as defined here is

different from that defined in Rosenberg et al. 2003 [25]). We designated $\delta_{study\text{-}AA\text{-}EA}$ as the measure of marker efficiency between EA and AA in our study populations, and $\delta_{reference\text{-}study\text{-}EA}$ or $\delta_{reference\text{-}study\text{-}AA}$ as the quantitative difference in efficiency between marker characteristics as they were reported previously, and as we observed them in the study populations. We observed that the maximum $\delta_{study\text{-}AA\text{-}EA}$ was 0.82, for the marker FY, and the minimum $\delta_{study\text{-}AA\text{-}EA}$ was 0.15. The mean was 0.32 and median was 0.28. Larger observed $\delta_{study\text{-}AA\text{-}EA}$ corresponded to greater marker efficiency for differentiating the EA and AA study populations. Furthermore, smaller values of the $\delta_{reference\text{-}study}$ (including $\delta_{reference\text{-}study\text{-}EA}$ or $\delta_{reference\text{-}study\text{-}AA}$) indicate that the marker as observed is more similar to the marker as described in the reference (and therefore the reported allele frequencies were relatively accurate for LBM training). For markers with higher values of this measure, since they did not match the training frequencies as well, their utility in practice was reduced. An efficient classification marker would be one with bigger $\delta_{study\text{-}AA\text{-}EA}$ and smaller $\delta_{reference\text{-}study}$ when the reference allele frequencies are used for training for the LBM. Figure 1 shows the relationship of these three δ measures; the straight line in the Figure 1(1) indicates the equality of $\delta_{study\text{-}AA\text{-}EA}$ and $\delta_{reference\text{-}study}$. Thus, Figure 1(1) illustrates that the majority of the markers have $\delta_{study\text{-}AA\text{-}EA} > \delta_{reference\text{-}study}$, and Figure 1(2) shows the ratio of $\delta_{reference\text{-}study\text{-}AA}$ to $\delta_{reference\text{-}study\text{-}EA}$ with a horizontal line specifying $\delta_{reference\text{-}study\text{-}AA} = \delta_{reference\text{-}study\text{-}EA}$. (Twenty-two of 36 markers studied (61%) are above the horizontal
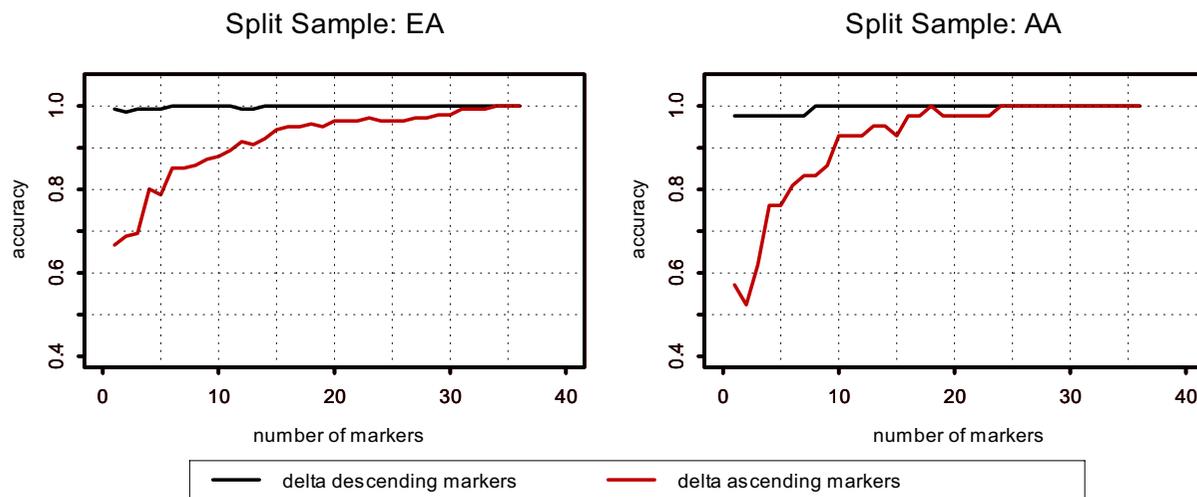
Figure 3
Assignment accuracy by LBM for Split samples. Assignment accuracy by LBM for Split samples. Split samples were used to evaluate the impact of reference group of allele frequencies on the assignment accuracy by LBM.

line, which indicates that they are less representative (of prior reports) for AAs than for EAs. This reduced correspondence of the observed AA allele frequency compared to the prior reports relative to our observations in EA populations, also causes decreased assignment accuracy in AAs compared to EA – in fact, the assignment accuracy in AAs never reaches 100%. Even with imperfect training frequencies, the LBM using the selected makers to classify individuals into subpopulations still performed very well, with average assignment accuracy of 96.8% and 99.9% for AA and EA respectively.) These results illustrate, further, that the selected marker panel is a relatively informative marker set in differentiating between EAs and AAs.

Assignment accuracy
In order to ascertain the smallest sufficient marker set and identify how many makers are needed to reach reasonable assignment accuracy, we took the approach of selecting markers by marker efficiency, as we did previously in evaluating the Bayesian method [1]. The relative assignment accuracy was evaluated by adding markers one-by-one up to 36 markers, with the order of δ either descending or ascending; the results are shown in Figure 2 (This result by LBM can be compared with results from STRUCTURE in Yang et al. 2005 [1]; cf. Figure 3, p. 308). FY was the most informative marker, and due to its unique value in distinguishing the EA and AA populations under study, we per-

formed analyses separately either including or excluding this marker.

In EAs (Figure 2, (1)), the assignment accuracy by LBM exceeded 99% using the most efficient marker FY, and reached 100% using the 10 most efficient markers excluding FY (when FY was excluded, the assignment accuracy using the next most efficient marker D11S936 dropped by 9%). In contrast, it would take 29 markers to reach >99% assignment accuracy when the least efficient markers are selected or the seven most efficient markers are omitted. In AAs (Figure 2, (2)), the assignment accuracy reached 96.4% using FY, and then the assignment accuracy changed inconsistently as more markers were added up to 21 markers, at which point assignment accuracy stabilized at 97.6%, achieving the maximum of 98.8% when all 36 markers were used. Overall, using LBM, it can exceed 95% when using at least the 9 most efficient markers. When FY was excluded, the assignment accuracy dropped by 38%.

This 38% drop, which reflects the difference in accuracy between the most efficient marker, FY, and the second most efficient one, D11S936, was further investigated by a corresponding analysis in which the study sample was randomly split into two groups and one group was treated as a reference sample. The drop declined to 6%, which was more comparable to the 9% in EAs. Thus, this reduced
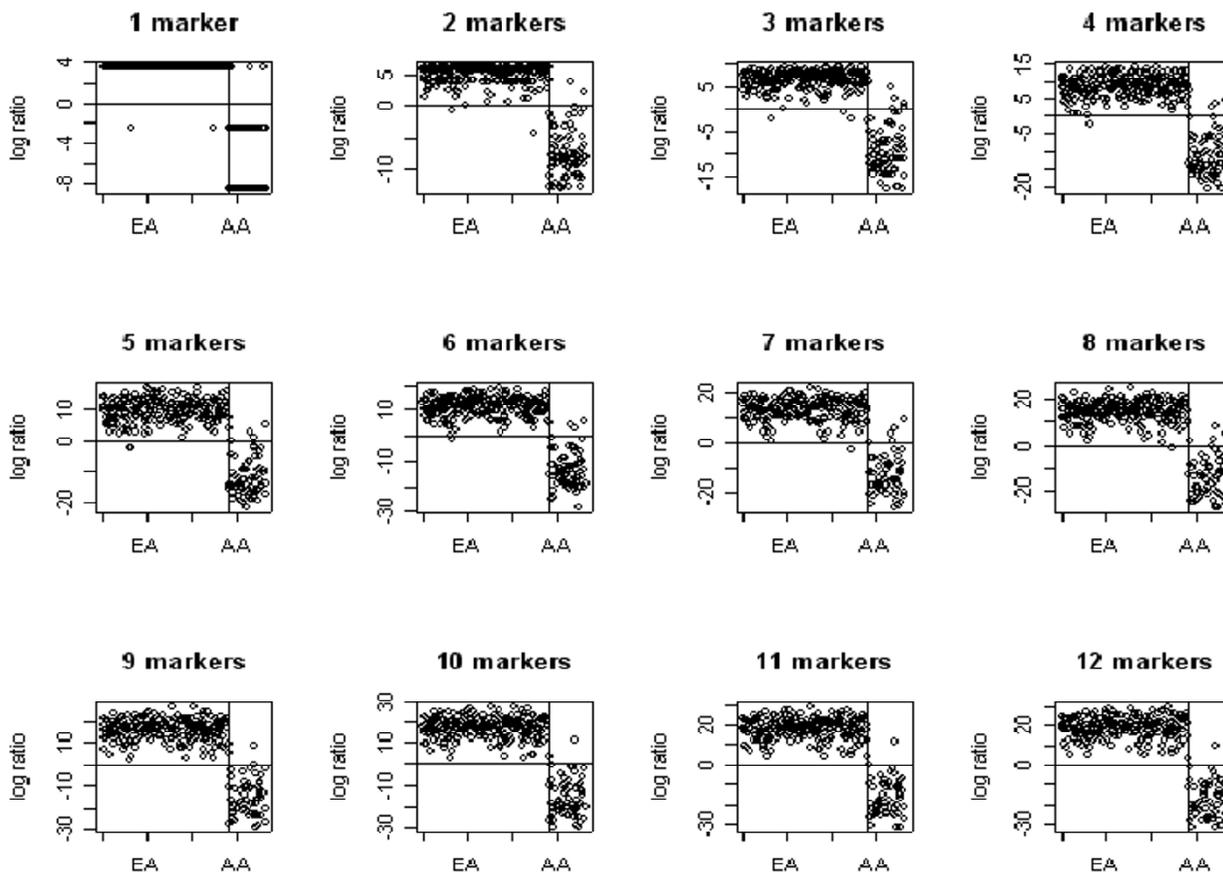
Figure 4
Logarithm of likelihood ratio for each individual. Logarithm of likelihood ratio for each individual grouping by their self-identified ethnicity. Markers were added one by one with δ descending. The first marker is FY.

accuracy was in large part attributable to mismatch between reported training allele frequencies and frequencies that are more representative of our Northeastern US AA population. LBM never reaches perfect assignment accuracy for AAs in this sample even when all the 36 markers were used, but accuracy did reach 98.8%.

Comparison of observed and reference allele frequencies
The high assignment accuracy by LBMs was observed notwithstanding the deviation between our observed allele frequencies and the reference frequencies described above. We further compared our observed allele frequencies with published reference allele frequencies using the χ^2 test. In EAs, after adjusting for sample size, there were 19 markers that differed at $p < 0.05$, while in AAs, the corresponding number of markers was 29. In other words,

allele frequencies observed in EAs matched the reference group more closely than did allele frequencies observed in AAs. As a result, the LBM performed better in EAs than AAs, as might be expected given the dependence of LBMs on prior knowledge of allele frequencies.

Evaluation of the influence of mismatched reference allele frequencies on assignment accuracy by means of split samples
As noted above, in many cases our observed allele frequencies showed nominally significant differences from population reference frequencies. This could reflect, for example, sampling error, or differences in allele frequency for population groups with similar self-identified ethnicity that are assessed at different geographic locations. To further assess the impact of the reference group on the
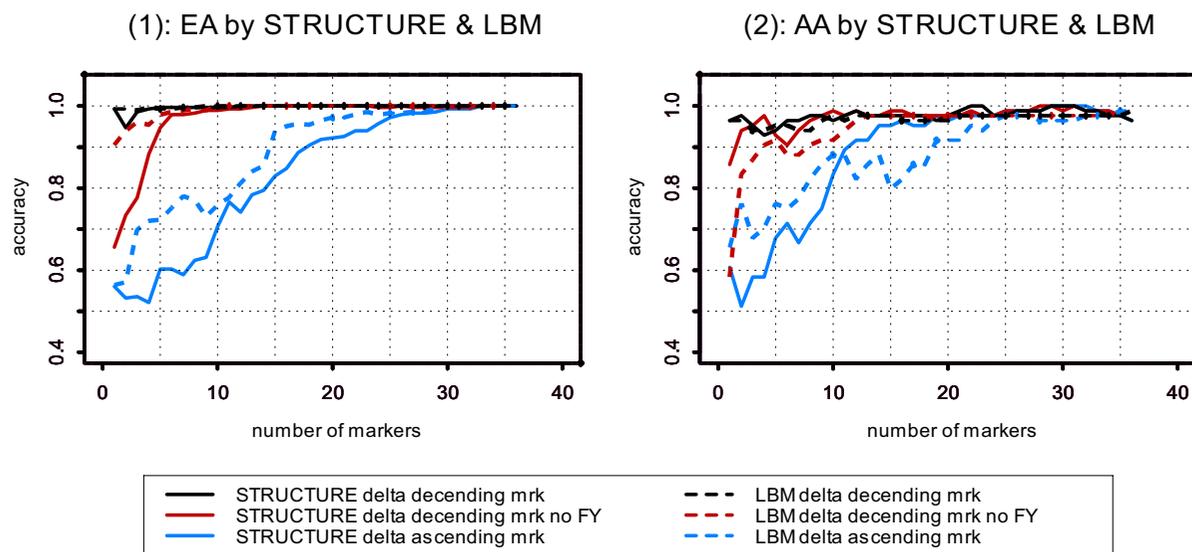
Figure 5
Comparison of STRUCTURE and LBM on assignment accuracy. The markers are adding one by one either by delta descending or ascending. Assignment accuracy without FY, the most efficient marker in the panel studied, was also evaluated.

assignment accuracy for LBM, we randomly split our EA and AA study datasets each into two equal-sized samples, treating one as the study group and the other as the reference group. Thus, we were able to model geographically appropriate allele frequencies for each group, at the expense of reducing the analysis sample size by a factor of two. The distributions of the allele frequencies for the two split samples are the same in EAs and AAs for all the markers based on the χ^2 test (p-value ranges from > 0.57 to 1). The results (Figure 3) for AAs using internal split samples improved dramatically compared to the results using the external reference group in AAs (Figure 2). These results (Figure 3) illustrate that the performance of the LBM depends greatly on how representative the reference allele frequencies are to those of the population being assigned when the parental population is known.

Logarithm likelihood ratio
We also calculated the logarithm of the likelihood ratio, expressing the comparison of the probability of being in the EA group compared to the AA group, based on formula (2) (Methods section), and generated a visual display of correct or misplaced group assignment for each individual, adding the markers one by one using a descending value of δ. Figure 4 shows the 12 most efficient markers. The horizontal line represents a log likelihood ratio of zero; those above zero are allocated to EA,

and below zero to AA (refer to equation (2)). The vertical line separates the groups. Therefore, those in the upper right and lower left quadrants are misclassified based on self-identified race. The first graph represents the allocation of each individual using only the most efficient marker, FY. As markers are added to the analyses, the log likelihood ratios increase and the separation between clusters become more and more marked. (Note that the Y-axis scale is not constant.)

One individual in the AA series appeared to be misclassified; see Figure 4 with 9 to 12 markers. Based on this observation, we examined the phenotypic information for this subject, and determined that, although self-identified as AA, the subject had one AA and one EA parent.

Comparison of LBM results with Bayesian results obtained using STRUCTURE
We compared the performance of LBM with results obtained using STRUCTURE and the same panel of markers by Yang et al. 2005 [1] (Figure 5); the samples used for Figure 5 are exactly the same as those for Figure 3 in Yang et al. 2005 [1] (cf. Figure 3, p. 308). In EAs (Figure 5 – (1)), the LBM provided more accurate group assignment than STRUCTURE, with the FY locus included or excluded. In AAs (Figure 5 – (2)), the relative performance of STRUCTURE and LBM was mixed.

Discussion

The LBM is appealing for population group assignment because it is straightforward and easily implemented, provided that sufficiently accurate reference allele frequencies are available. We provide a set of allele frequencies for all markers herein that will prove useful for classifying populations similar to those discussed in the present article [see Additional file 1]. Under these circumstances, the LBM should classify individuals at least as accurately as STRUCTURE, and probably more accurately. However, a representative reference population may be difficult to establish in some cases. With a good reference group, as shown in the analysis of split samples (p-values of the χ^2 test range from >0.57 to 1 for distributions of allele frequencies for the two split samples), LBM performed very well. In EAs, the clustering by LBM is as good as by STRUCTURE (using an ancestry model of admixture and an allele frequency dependence model) for δ descending, but LBM performs better for ascending values of δ. For AAs, LBM and STRUCTURE cluster the groups equally well. STRUCTURE retains certain advantages, such as the ability to classify individuals by proportional ancestry for subsequent application of the structured association method, as discussed elsewhere [1]. It should be noted that the superior performance of LBMs over STRUCTURE, when observed, depends on LBM having more data available than STRUCTURE in the form of reference allele frequencies.

The observed allele frequencies in this study matched reference allele frequencies better for EA than AA populations. Subjects from some populations from different geographic areas might have quite different admixture proportions and ancestral origins. This is demonstrably the case for African-Americans, since in different parts of the US the percent admixture from EAs is known to range at least from 12% to 23% [26]. Another issue with LBM involves justification for the multiplication of allele frequencies across loci under the assumption of linkage equilibrium. If the allele frequencies of different STRs vary among subpopulations, then the loci are not in complete linkage equilibrium or are not statistically independent even if they are genetically unlinked. However, we did assume linkage equilibrium within the subpopulations. This is also an underlying assumption for STRUCTURE [4]. This assumption might prove to be problematic under some circumstances, but the practical impact seemed minimal for the present study, as evidenced by the fact that LBM performed well.

The result from the single most informative marker, FY, could exceed 99% and 96% assignment accuracy in EAs and AAs, respectively. This result is, of course, sample-specific to some extent; AA subjects who are homozygous for the allele more characteristic of European ancestry (i.e., FY

(+/+)), should have a population frequency of about 4%, given a 20% admixture rate from EA, and would be misclassified into the EA group if based only on this marker; this misclassification rate is equal to what we observed, about 4% in AAs. Likewise, EAs heterozygous for the FY(-) allele characteristic of AAs are observed as well, and they are liable to be misclassified as AAs. Our Northeastern US AA population had approximated the expected European admixture rate, based on the information from FY.

The sample size of the populations being assigned is not an issue for LBM, while it is for STRUCTURE. The Bayesian cluster approach taken by STRUCTURE requires building a likelihood function from the observed samples to infer allele frequencies, such that if the sample size is insufficient, the estimated allele frequencies might not be accurate. As a result, sample size in each subgroup affects the assignment accuracy, and our simulation result [1] shows that approximately fifty subjects are required to have stable assignment accuracy by STRUCTURE. LBM, in contrast, uses allele frequencies from the reference populations; there is no need to estimate allele frequencies by LBM. Thus, even a single individual can be assigned accurately using the LBM.

We conclude that assignment accuracy by LBM depends on the efficiency of the markers selected (FY alone can separate EAs and AAs with accuracy that can approach 99% for excluding AAs from a presumed EA sample), the number of markers (other things being equal, more markers produce higher assignment accuracy), and greatly on how representative the parental population reference allele frequencies are for the populations being queried.

Methods
Subjects
Three hundred sixty-six individuals recruited in the Northeastern US (classified as 282 EAs, 84 AAs) were studied. These individuals were selected from a larger sample for evaluation of this likelihood based method because they had complete marker data for all markers described below. All subjects provided informed consent as approved by the appropriate institutional review boards.

Markers genotyped
Detailed marker and genotyping information was described previously [1]. Briefly, two different sets of STR markers were used. First, we used the AmpFLSTR Identifiler PCR Amplification Kit (Applied Biosystems (ABI), Foster City, CA), which provides data from a set of 16 loci useful for forensic purposes (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S443, vWA, TPOX, D18S51, D5S818, FGA, and amelogenin). Amelogenin is used for sex identification rather than for polymorphism content, so informa-

tion from that locus was not included in any analyses. Second, we selected 21 markers known to have high δ between EAs and AAs, and in some cases Hispanic and Asian populations, based on the report of Smith et al. 2001 [27]. This marker panel includes markers D1S196, D1S2628, D2S162, D2S319, D5S407, D5S410, D6S1610, D7S640, D7S657, D8S272, D8S1827, D9S175, D10S197, D10S1786, D11S935, D12S352, D14S68, D15S1002, D16S3017, D17S799, and D22S274. We also genotyped marker FY, added to the 36 STRs because of its known value in identifying individuals of primarily African ancestry.

Measures of marker efficiency
δ was used to measure the marker efficiency. The definition and properties of δ are described in Yang et al. 2005 [1]. Briefly, δ is half the sum of the absolute difference in population frequency over all alleles for each marker between two populations.

Analysis with the likelihood-based method (LBM)
We assumed HWE among alleles for each marker within populations and linkage equilibrium between markers. The likelihood, or the probability of observed genotype profile, for each individual to be in a specific population is calculated as

$$L_Z = \Pr(X \mid Z, \ P_Z(p_{11}, \ p_{12}, \ldots, p_{mn_m})) = \prod_m (p_{mi}^2)^h \times (2p_{mi}p_{mj})^{1-h} \qquad (1)$$

where X is a vector of genotypes of marker loci, Z is the proposed population of origin, $P_Z(p_{11}, \ p_{12}, \ldots, \ p_{mn})$ is the set of reference allele frequencies $p_{11}, \ p_{12}, \ldots, \ p_{mn_m}$ for the n_m alleles of m markers of population Z, and n is a dummy variable for homozygosity (i.e., when the locus is homozygous, h is 1, otherwise h is 0) for each marker locus. When an allele is absent for a given population in the reference frequencies, the corresponding allele frequency in the study group is estimated and used in the calculation of likelihood. An individual is assigned to a population if the maximum likelihood results from assignment to that population among all possible population-specific likelihoods calculated. For assigning individuals into one of two populations A or B, an individual is assigned to population A if the logarithm of likelihood ratio is greater than zero, or otherwise to B, as shown in equation (2).

$$\log\left(\frac{L_{Z=A}}{L_{Z=B}}\right) = \frac{\Pr(X \mid Z = A, P_A(p_{11}, p_{12}, \ldots, \ p_{mn_m}))}{\Pr(X \mid Z = B, P_B(p_{11}, p_{12}, \ldots, \ p_{mn_m}))} > 0. \qquad (2)$$

An individual was considered to be assigned accurately to a group when the greatest likelihood among all the calcu-

lated likelihoods assigned an individual the same ethnicity as the self-identified population group of that individual. Assignment accuracy in each population group was defined as the proportion of correctly assigned ethnicities. (The above decision rule is optimal if we have equal priors of proportion for the two ethnic groups. However, when there are more people from one group, *a priori*, then the prior information needs to be incorporated to improve the overall performance in terms of misclassification rate.) The method was realized in R/S-Plus; the function codes are available upon request from the authors.

The initial set of reference population-specific allele frequencies (training frequencies) for the 36 markers were derived from ABI reference materials [27] or Smith et al. 2001 [28], depending on the source of each marker. Since ABI uses different nomenclature (in some cases) and we redesigned some primers referenced by Smith to facilitate efficient genotyping, each observed allele had to be matched to the corresponding allele for each marker. Alleles at one marker locus (D6S1610) described by Smith et al. 2001 [28] could not be matched accurately to our data from the same marker (however, the value of EA/AA δ that we derived for that marker, 0.336, was similar to the value reported by Smith et al., which was 0.337). The χ^2 test was used to compare the allele distributions of the study group and the reference group.

Evaluation of the impact of the training frequencies on population group assignment accuracy
To evaluate the impact of the training frequencies on population group assignment accuracy, we compared the literature-derived training allele frequencies (described above) with training allele frequencies computed from our specific populations. To do so, we randomly split the 282 EAs and 84 AAs into two equal-sized groups. One was treated as the study group, and the other was treated as the reference group, from which the training allele frequencies were estimated.

Authors' contributions
BZY participated in study design, wrote the computer code, carried out the statistical analyses and drafted the manuscript. HZ participated in study design, and conceptual and technical assistance. HRK provided sample recruitment and phenotyping, and commented on the manuscript. JG designed the study, coordinated genotyping efforts and helped draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Population allele frequencies in the EA and AA samples for the 35 STRs markers and the FY marker studied. For each marker, the marker name and alleles are listed with the allele frequencies.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2156-6-50-S1.doc]

References

1. Yang BZ, Zhao H, Kranzler H, Gelernter J: Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. *Genetic Epidemiology* 2005, 28:302-312.
2. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997-1004.
3. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**:945-959.
4. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations.** *Am J Hum Genet* 2000, **67**:170-181.
5. Reich DE, Goldstein DB: **Detecting association in a case-control study while correcting for population stratification.** *Genet Epidemiology* 2001, **20**:4-16.
6. Ripatti S, Pitkaniemi J, Sillanpaa MJ: **Joint modeling of genetic association and population stratification using latent class models.** *Genet Epidemiology* 2001:S409-14.
7. Satten GA, Flanders WD, Yang Q: **Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model.** *Am J Hum Genet* 2001, **68**:466-477.
8. Sillanpaa MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P: **Bayesian association mapping for quantitative traits in a mixture of two populations.** *Genetic Epidemiology* 2001:S692-9.
9. Zhang S, Zhao H: **Quantitative similarity-based association test using population samples.** *Am J Hum Genet* 2001, **69**:601-614.
10. Pfaff C, Kittles R, Shriver MD: **Adjusting for population structure in admixed populations.** *Genetic Epidemiology* 2002, **22**:196-198.
11. Zhang S, Zhu X, Zhao H: **On a semi-parametric test to detect associations between quantitative traits and candidate genes using unrelated individuals.** *Genetic Epidemiology* 2003, **24**:44-56.
12. Chen HS, Zhu X, Zhao H, Zhang S: **Qualitative semi-parametric test to detect genetic association in case-control design under structured population.** *Annals of Human Genetics* 2003, **67**:250-264.
13. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies.** *Genetics* 2003, **164**:1567-1587.
14. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Clayton DG, McKeigue PM: **Control of confounding of genetic associations in stratified populations.** *Am J Hum Genet* 2003, **72**:1492-1504.
15. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D: **Methods for high-density admixture mapping of disease genes.** *Am J Hum Genet* 2004, **74**:979-1000.
16. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De The G, Essex M, Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D: **A high-density admixture map for disease gene discovery in African Americans.** *Am J Hum Genet* 2004, **74**:1001-13.
17. Tang H, Peng J, Wang P, Risch NJ: **Estimation of individual admixture: Analytical and study design considerations.** *Genetic Epidemiology* 2005, **28**:289-301.
18. Paetkau D, Calvert W, Sterling I, Strobeck C: **Microsatellite analysis of population structure in Canadian polar bears.** *Molecular Ecology* 1995, **4**:347-354.
19. Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C: **An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations.** *Genetics* 1997, **147**:1943-1957.
20. Rannala B, Mountain JL: **Detecting immigration by using multilocus genotypes.** *Proc Natl Acad Sci USA* 1997, **94**:9197-9201.
21. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M: **New methods employing multilocus genotypes to select or exclude populations as origins of individuals.** *Genetics* 1999, **153**:1989-2000.
22. Guinand B, Topchy A, Page KS, Burnham-Curtis MK, Punch WF, Scribner KT: **Comparison of likelihood and machine learning methods of individual classification.** *The Journal of Heredity* 2002, **93**:260-269.
23. Manel S, Berthier P, Luikart G: **Detecting wildlife poaching: Identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes.** *Conservation Biology* 2002, **16**:650-659.
24. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A: **GENECLASS2: a software for genetic assignment and first-generation migrant detection.** *J Hered* 2004, **95**:536-9.
25. Rosenberg NA, Li LM, Ward R, Pritchard JK: **Informativeness of genetic markers for inference of ancestry.** *Am J Hum Genet* 2003, **73**:1402-22.
26. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD: **Estimating African American admixture proportions by use of population-specific alleles.** *Am J Hum Genet* 1998, **63**:1839-1851.
27. **Website title** [http://home.appliedbiosystems.com/]
28. Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, O'Brien SJ: **Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations.** *Am J Hum Genet* 2001, **69**:1080-94.