BMC Bioinformatics

BioMed Central

Software

PSMIX: an R package for population structure inference via maximum likelihood method

Baolin Wu[1], Nianjun Liu[2] and Hongyu Zhao*[3,4]

Address: [1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA, [2]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA, [3]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, Yale University School of Medicine, New Haven, CT, USA and [4]Department of Genetics, Yale University School of Medicine, New Haven, CT, Yale University School of Medicine, New Haven, CT, USA

Email: Baolin Wu - baolin@biostat.umn.edu; Nianjun Liu - nliu@uab.edu; Hongyu Zhao* - hongyu.zhao@yale.edu

* Corresponding author

Published: 22 June 2006

BMC Bioinformatics 2006, **7**:317 doi:10.1186/1471-2105-7-317

Received: 23 January 2006
Accepted: 22 June 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/317

Abstract

Background: Inference of population stratification and individual admixture from genetic markers is an integrative part of a study in diverse situations, such as association mapping and evolutionary studies. Bayesian methods have been proposed for population stratification and admixture inference using multilocus genotypes and widely used in practice. However, these Bayesian methods demand intensive computation resources and may run into convergence problem in Markov Chain Monte Carlo based posterior samplings.

Results: We have developed PSMIX, an R package based on maximum likelihood method using expectation-maximization algorithm, for inference of population stratification and individual admixture.

Conclusion: Compared with software based on Bayesian methods (e.g., STRUCTURE), PSMIX has similar accuracy, but more efficient computations.

PSMIX and its supplemental documents are freely available at http://bioinformatics.med.yale.edu/PSMIX.

Background

Information about population structure, namely population stratification and admixture, is useful in a variety of situations, such as association studies of genes underlying complex traits, subspecies classification, genetic barrier detection, and evolutionary study [1-10]. For example, it is very important to identify genetic ancestry and admixture in admixture mapping [7,8]. The presence of population stratification or admixture may pose a practical nuisance as well. In association studies, case-control design is often used to identify genetic variants underlying complex traits by comparing allele frequencies between unrelated individuals that are affected and those unaffected. However, the presence of population stratification or admixture in the sample can lead to spurious associations between a candidate marker and a phenotype [5,10,11]. In forensic studies, the identification of reference groups is central but becomes difficult when there exists population stratification [12,13]. In the estimation of the magnitude of inbreeding, it is useful to distinguish between the causes for the excess homozygosity which might be consanguineous mating or population substructure, or an artifact due to factors like null alleles [14]. In all these situations, identifying population stratification or admixture has been an important component.

Population structure can be identified based on visible characters such as language, culture, physical appearance, and geographic region. But this can be subjective and may bear no relevance to genetics. Evanno et al. [15] gave a good example by mentioning migratory bats which can be found thousands of kilometers apart but from the same breeding roost in winter [16]. Statistical methods have been proposed to infer population stratification and individual admixture using multilocus genotype data [1,2,17-31]. Different methods use different statistical tools and population genetic assumptions. Pritchard et al. [2] introduced a model-based clustering method to infer population structure and assign individuals to populations using multilocus genotype data. They used Bayesian formulation and generated the posterior distributions using a Markov Chain Monte Carlo (MCMC) method based on Gibbs sampling. Their main modeling assumptions are Hardy-Weinberg equilibrium (HWE) within populations and linkage equilibrium (LE) between markers within each population. Falush et al. [21] extended the method to allow for loose linkage between loci. The method of Corander et al. [17,18] uses multilocus molecular markers and geographical information provided by the sampling design. Unlike the methods of Pritchard et al. [2] and Falush et al. [21], the methods of both Dawson and Belkhir [19] and Corander et al. [17,18] can directly estimate the number of (sub)populations and assign individuals to the (sub)populations. The main difference between the two approaches is the parametric assumption of the number of populations [17-19,32]. Corander et al. [18] considered the geographical sampling design of the individuals and set the maximum number of populations allowed to be the number of locations sampled, whereas for Dawson and Belkhir [19], it is the total number of individuals. Corander et al. [17] generalized the approach of Corander et al. [18] and it became more similar to the approach of Dawson and Belkhir [19] in terms of model assumptions and some technical details, especially when the data is specified for individual level analysis. Guillot et al. [23,24] used spatial statistical models employing both landscape ecology and population genetics information, which is especially useful in situations of young populations exhibiting low genetic differentiation [23,33]. Excoffier et al. [20] applied approximate Bayesian computation method to the estimation of all the parameters of an explicit admixture model. Their method can easily deal with complex mutation models and partially linked loci and is superior when the admixture is more ancient [20]. The majority of the methods for population structure inference are Bayesian approaches [1,2,17-26] with few exceptions such as Tang et al. [30], Satten et al. [29], Wang [31], and Purcell and Sham [28]. Meanwhile, several methods have been proposed for the assignment of individuals to populations [34-36]. As for computer programs available based on existing methods, the majority are also based on Bayesian MCMC methods, such as STRUCTURE [2,21], GENELAND [24], BAPS/BAPS 2 [17,18], and ADMIXMAP [25,37,38], with the exception of L-POP[28] which is based on latent class analysis. Table 1 summarizes some of the commonly used software for population structure inference. STRUCTURE is the most commonly used program for population structure inference which has been used both on humans [4,13,39] and other species [3,40-42] (at the time this article is written, the paper of Pritchard et al. [2], where the method of STRUCTURE was originally proposed, has been cited about 760 times). We choose to compare the performance of our package with STRUCTURE and L-POP (the representative of the frequentist methods).

Implementation

We have developed an efficient R package, named PSMIX (Population Structure inference via MIXture model), for population stratification and individual admixture inference. Since R can be slow when computation is intensive, we implemented the expectation-maximization (EM) algorithm [43] using C programming language. PSMIX is mainly based on the methods proposed in Tang et al. [30] and Liu et al. [27]. Three models (described in section 2.2, 2.3, and 2.4, respectively) are discussed in full detail in [27]. The second one is equivalent to the model proposed in Tang et al. [30]. The first model is a special case of the second one. In Tang et al. the method itself has been fully assessed by simulation studies [30].

Results

We used two real datasets from Rosenberg et al. [4] and one simulated dataset from Tang et al. [30] to demonstrate the functionality of PSMIX. One real dataset contains two American populations, Pima and Surui with 25 and 21 individuals, respectively; the other contains two European populations, Sardinian and Russian with 28 and 25 individuals, respectively. The simulated data set contains 50 individuals from each of the two ancestral populations, and 200 individuals from the admixed population. The true individual admixture values of the admixed individuals are also available.

To evaluate the efficiency of PSMIX, we randomly selected 100 markers from the Pima-Surui dataset with no missing values and tried the four models available in STRUCTURE2.0. Burnin length and number of MCMC replications after burnin were both set to be 10,000 in the analyses. The time needed for each run of STRUCTURE2.0 increased almost linearly with the increase of number of clusters. On our PC with Pentium III 500 MHZ CPU and 384 MB SDRAM, when K = 2, about two and a half min-

Table 1: Commonly used software for population structure inference

Software	STRUCTURE	GENELAND	BAPS/BAPS2	ADMIXMAP	PARTITION	L-POP	PSMIX
Method	Bayesian MCMC	Bayesian MCMC	Bayesian, MCMC is used when the number of populations ≥ 9	Bayesian MCMC	Bayesian MCMC	Latent class analysis, EM	Clustering analysis, EM
Features	Population structure inference	Process geo-referenced individual multilocus genetic data for population structure inference	Population structure inference. Use geographical sampling design of the individuals	Mainly for analysis of datasets that consist of trait measurements and genotype data on a sample of individuals from an admixed or stratified population	Population structure inference	Population structure inference	Population structure inference
Assumptions	HWE and LE between loci	HWE, LE between loci, and spatial distribution of sub-populations	HWE and LE between loci	Ancestry state is the same at all loci within a compound locus on any gamete. Mating is not assortative for admixture in the population from which the parental gametes were drawn	HWE and LE between loci. The underlying population genetic model is appropriate for out-crossing diploid organisms.	HWE and LE between loci	HWE and LE between loci
Input parameters	Parameters for running MCMC, parameters for ancestry model and allele frequency model, and the number of populations	In addition to genetic and spatial data, the user must provide parameters for the maximum number of populations, the way geographical information is handled and the allele frequency model	When MCMC is used, need parameters for running MCMC	Parameters for running MCMC, allele frequencies (number of population is specified here), and mating model. Disease information (outcome variable is suggested even if focus on population structure). Parameters for tests and output	Parameters for running MCMC, maximum number of populations, prior parameter for allelic diversity, and prior parameter for number of populations	Number of populations, admixture option, data format options, model options, output format options, and convergence criterion	Number of populations and convergence criterion
Output	One file for estimates and some files for plots. Main parameter estimates are inferred ancestry of individuals and estimated allele frequencies in each population	Main parameter estimates are the number of populations, population membership of each individual, maps giving the population memberships of each geographical pixel of a given size to locate genetic discontinuities between populations	Main parameter estimates are the number of populations and population membership of each individual	Individual/gamete level admixture variables. Ancestry-specific allele or haplotype frequencies. Results for association analysis and model parameters	The output file contains a list of the parameter settings followed by the sequence of observations of the Markov chain. A companion program PartitionView is provided to obtain useful information from the PARTITION output file.	Main outputs are estimates of allele frequencies, posterior class probabilities, and class-specific allele frequencies	Main parameter estimates are inferred ancestry of individuals
Advantages	Easy to use. Once number of populations is given, the estimates are accurate	Easy to use. Flexible to extend. Can work with or without spatial information. Can estimate number of populations	Easy to use. Provide good estimate for number of populations. When geographical sampling information is applicable, can improve the statistical power to detect clusters in the data	In addition to population structure inference, can perform association analysis on structured populations. Can deal with tightly linked loci using haplotypes	Easy to use. Can estimate number of populations and calculate a Bayes factor in support of a single source population against the alternative of more than one source population.	Computationally efficient	Easy to use. Computationally efficient. Flexible to extend

Table 1: Commonly used software for population structure inference *(Continued)*

Limitations	Computationally intensive. Can detect number of populations but does not work well.#	Does not handle admixture. Computationally intensive, especially when "Falush" is used as allele frequency model, or number of populations needs to be estimated.	Very memory intensive. When MCMC is used, becomes relatively computationally intensive. Only provides membership partition, does not handle admixture	Difficult to use. Computationally intensive. Does not estimate number of populations *	Computationally intensive, especially when number of populations needs to be estimated	Parameter configuration is difficult to use, works OK for discrete populations but not for admixed populations. Does not estimate number of populations	Does not estimate number of populations
Platforms	Windows, Unix/Linux	Windows/Linux/Mac (R package)	Windows	Windows, Unix/Linux. R statistical package is required	Windows	Windows (DOS), Unix/Linux	Windows/ Linux/Mac (R package)
References	Pritchard et al. (2000), Falush et al. (2003)	Guillot et al. (2005)	Corander et al. (2003, 2004)	McKeigue et al. (2000), Hoggart et al. (2003, 2004)	Dawson and Belkhir (2001)	Purcell and Sham (2004)	Tang et al. (2005), Liu et al. (2005)
URL	http://pritch.bsd.uchicago.edu/structure.html	http://www.inapg.inra.fr/ens_rech/mathinfo/personnel/guillot/Geneland.html	http://www.rni.helsinki.fi/~mjs	http://www.lshtm.ac.uk/ncdeu/genetics/#admix	http://www.genetix.univ-montp2.fr/partition/partition.htm#analyse_+exe	http://statgen.iop.kcl.ac.uk/lpop	http://bioinformatics.med.yale.edu/PSMIX

# With the findings of Evanno et al. (2005), STRUCTURE's ability to detect number of populations should be improved greatly.
* Only focus on the function of population structure inference.

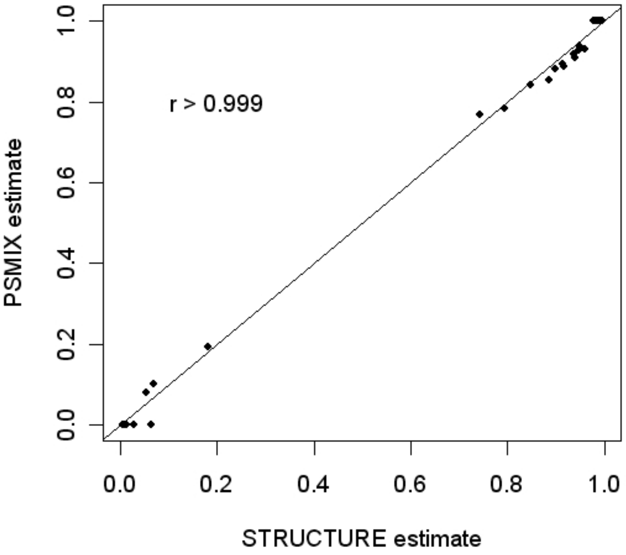

Figure 1
Comparison of estimates of individual admixture of STRUCTURE (x-axis) and PSMIX (y-axis) for the data of Pima and Surui from Rosenberg et al. (2002). Only the first 50 markers were used.

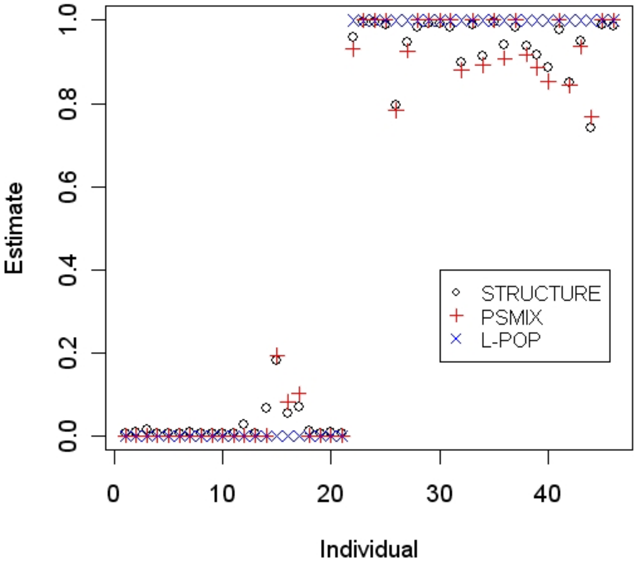

Figure 2
Estimates of STRUCTURE, PSMIX and L-POP for the data of Pima and Surui (the first 21 samples are Surui) from Rosenberg et al. (2002). Only the first 50 markers were used.

utes were needed for each run of STRUCTURE2.0. For all PSMIX runs, we set the stopping criterion to be that the parameter difference $<10^{-6}$ between consecutive iterations, or 10,000 steps, whichever was reached first. For the same Pima-Surui data with 100 markers, each run of PSMIX needed about 6 seconds.

To evaluate the accuracy of PSMIX, we compared the results of PSMIX with those of STRUCTURE. Figure 1 gives a sample run for the Pima-Surui dataset using only the first 50 markers. For the Pima-Surui dataset using the first 50 markers from the original data, the correlation coefficient between the results of PSMIX and STRUCTURE was greater than 0.999. For the Sardinian-Russian dataset, when all 377 markers were used, both STRUCTURE 2.0 (use independent allele frequencies among populations with admixture model) and PSMIX had one individual misclustered. The correlation coefficient between the results was 0.906. The two methods produced very similar results. This is consistent with the findings in Tang et al. [30]. Figure 2 shows the results of STRUCTURE, L-POP, and PSMIX. We can see that the results of PSMIX are much closer to those of STRUCTURE.

To evaluate the performance of PSMIX, we also used a simulated data set exhibiting population admixture. From Figure 3 we can see the PSMIX performs pretty well and the results are almost identical to those from STRUCTURE.

Discussion

We have implemented a likelihood based method of population structure inference into an efficient R package, PSMIX. PSMIX can be used in population genetics and disease gene mapping, wherever population stratification or individual admixture is needed to be estimated from genetic markers. Compared with other available similar programs, PSMIX has several advantages. First, it is computationally efficient and provides similar accuracy under realistic situations (Tang, et al. [30] and Liu et al., Technical Report [27]). And thus the confidence intervals of the estimates can be constructed via resampling methods, e.g., the bootstrap method [30]. Second, as shown in Tang et al. [30], it performs a little better (compared with STRUCTURE) under some conditions involving a small number of ancestors and markers. We note that L-POP is also computationally efficient. However, it is not clear if L-POP can perform better under such conditions. Third, it is very flexible. It is likelihood based and can be easily incorporated into study designs, such as marker choice [30]. The program is implemented as a public R package and can be easily extended and incorporated into other packages. This is an advantage of PSMIX over STRUCTURE and L-POP, which has only executable programs.

We would like to note that the examples used in this work are mainly for the purpose of demonstrating the R package, not for the purpose of the assessing the underlying method. Please refer to Tang et al. [30] for a detailed assessment of the methodology.

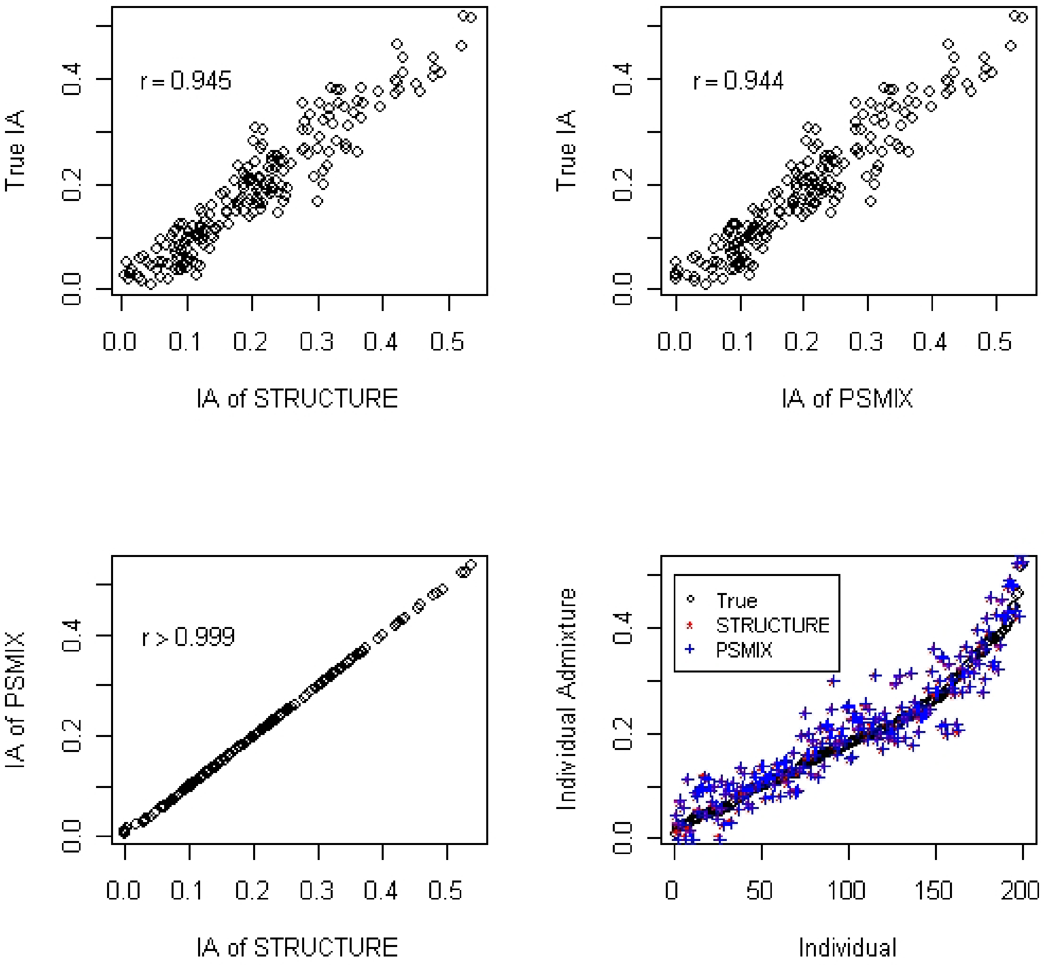

Figure 3
Estimates of STRUCTURE and PSMIX for the simulated data.

In our simulation and application to real data, PSMIX and STRUCTURE gave very similar results. This is not surprising because estimating parameters via maximum likelihood and maximum a posterior with flat prior is formally strictly similar, where PSMIX belongs to the former and STRUCTURE belongs to the latter.

Many studies have been performed to assess the ability of STRUCTURE in assigning individuals to their populations of origin using either real data or simulated data [3,44-47]. However, very limited studies have been performed to assess the ability of STRUCTURE in detecting the number of populations. Recently, Evanno et al. [15] performed a systematic study on this issue using simulations. They simulated amplified fragment length polymorphism (AFLP) and microsatellite genetic data under three population structure models: the island model, a contact zone, and a hierarchical island model [15]. Their major finding is that the "log probability of data", an ad hoc criterion suggested by Pritchard et al. [2] for detecting the number of populations, does not provide a correct estimation of the number of populations most of the time [15]. However, they found that another ad hoc statistic, which is based on the rate of change in the log probability of the data between successive numbers of populations, can accurately detect the uppermost hierarchical level of structure [15]. They also found some other factors that can affect the detection of the number of populations [15]. These findings are important and useful in that with the increasing usage of STRUCTURE, they provide guidance on how to use STRUCTURE to detect the number of populations. However, PSMIX does not directly detect the number of populations in this version. Due to its computation efficiency, model selection methods such as Akaike information criterion (AIC) [28,48], Bayesian information criterion (BIC) [49], and even more general, penalized likelihood based methods [50,51] can be used for this purpose. The findings of Evanno et al. [15] may be incorporated into PSMIX as well. This is one of our future works.

EM approach and Bayesian MCMC approach have their own advantages and disadvantages. They both can trap in local modes, although theoretically speaking, Bayesian MCMC approach can converge to the true value eventually, maybe after an unrealistic long time. However, the Bayesian MCMC approach, in addition, has the label switching problem. Two authors (Stephens and Donnelly) of the paper where the method of STRUCTURE was proposed [2] mentioned in other papers [52,53] on methods to deal with this problem. Although this issue is believed to be well addressed by STRUCTURE, it does make the Bayesian MCMC approach more complicated. However, this topic is beyond the scope of this work. From the users' point of view, they only see the computation efficiency and stability of the methods.

We think that it may be necessary to explicitly explain some details about the models mentioned in this work. First, the orientation of Tang et al. [30] is different from that of STRUCTURE, L-POP, and Liu et al. [27]. The goal of the former was to estimate individual admixture for the admixed individuals. The original focus of the latter was to "identify discrete clusters roughly corresponding to subpopulations" [30]. STRUCTURE, L-POP, and Liu et al. [27] use methods for clustering, although they "can also be applied to an admixture model" [30]. So initially, Tang et al. [30] faced a population (the "admixed group" in their paper) that is currently in Hardy-Weinberg equilibrium, but was created as the result of admixture at some point in the past. However, as emphasized in Tang et al. [30], the problem may not be identifiable without the inclusion of pseudo-ancestors who are proxies of the true pure ancestry [25]. Here the nonidentifiablity issue is related to the problem, and by no means pertains to the method. In other words, the nonidentifiablity issue exists and has nothing to do with the statistical methods to be used, if pseudo-ancestors are not included. Therefore, the actual data Tang et al. [30] dealt with consist of "I^0 individuals from the admixed group, as well as I^K subjects from each of the K ancestral populations" [30], that is, a stratified "pooled" population. So the actual data all these methods deal with are the same in the sense that the data consist of stratified populations within which Hardy-Weinberg equilibrium holds. One major difference is that Tang et al. [30] only focus on the individual admixture of the people in the admixed population (their original population). Facing the same data, the method in Tang et al. [30] is for clustering as well, in spirit. They included pseudo-ancestors and used clustering method in order to estimate individual admixture. In other words, all the aforementioned methods are for population stratification, and can be applied to estimate individual admixture. Thus the comparisons made in this work are appropriate. We also want to emphasize here the importance of inclusion of ancestral populations or their surrogates when individual admixture is needed; otherwise the problem may not be identifiable no matter what method to use.

Conclusion

In summary, we have implemented a new, likelihood based method for inference of population stratification and individual admixture which is available as a public R package. Although the package has several advantages over its peers, we strongly suggest that the users use different software in their analysis. If the results from these software are consistent; this may provide more support for the results; if the results are not consistent, further investigation is needed. A potential limitation is the assumption of independence among markers behind PSMIX, which will be addressed in future versions of PSMIX.

Availability and requirements

Project name: PSMIX

Project home page: http://bioinformatics.med.yale.edu/PSMIX

Operating system(s): MS Windows, Linux, Mac

Programming language: C, R

Other requirements: R 2.0 or higher

License: GPL

Any restrictions to use by non-academics: none

Authors' contributions

BW participated in the design of the study, implemented PSMIX, and helped to draft the manuscript; NL participated in the design of the study, performed the analysis, and drafted the manuscript; HZ conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the two anonymous reviewers for their insightful and constructive comments which have greatly improved the presentation of our work. We thank Dr. Hua Tang for sending us her presentation slides. This work was supported in part by NIH grants GM59507 and GM57672.

References

1. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly M, Reich D: **Methods for high-density admixture mapping of disease genes.** *Am J Hum Genet* 2004, **74(5):**979-1000.
2. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155(2):**945-959.

3. Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Maki-Tanila A, Tixier-Boichard M, Vignal A, Wimmersh K, Weigend S: **Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds.** *Genetics* 2001, **159(2):**699-713.
4. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations.** *Science* 2002, **298(5602):**2381-2385.
5. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association.** *Lancet* 2003, **361(9357):**598-604.
6. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D: **Assessing the impact of population stratification on genetic association studies.** *Nat Genet* 2004, **36(4):**388-393.
7. Montana G, Pritchard JK: **Statistical tests for admixture mapping with case-control and cases-only data.** *Am J Hum Genet* 2004, **75(5):**771-789.
8. Reich D, Patterson N: **Will admixture mapping work to find disease genes?** *Philos Trans R Soc Lond B Biol Sci* 2005, **360(1460):**1605-1607.
9. Marchini J, Cardon LR, Phillips MS, Donnelly P: **The effects of human population structure on large genetic association studies.** *Nat Genet* 2004, **36(5):**512-517.
10. Chen HS, Zhu X, Zhao H, Zhang S: **Qualitative semi-parametric test for genetic associations in case-control designs under structured populations.** *Ann Hum Genet* 2003, **67(Pt 3):**250-264.
11. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405(6788):**847-856.
12. National research council: **The Evaluation of Forensic DNA Evidence.** 1996.
13. Kim JJ, Verdu P, Pakstis AJ, Speed WC, Kidd JR, Kidd KK: **Use of autosomal loci for clustering individuals and populations of East Asian origin.** *Hum Genet* 2005, **117(6):**511-519.
14. Overall AD, Nichols RA: **A method for distinguishing consanguinity and population substructure using multilocus genotype data.** *Mol Biol Evol* 2001, **18(11):**2048-2056.
15. Evanno G, Regnaut S, Goudet J: **Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.** *Mol Ecol* 2005, **14(8):**2611-2620.
16. Petit E, Balloux F, Goudet J: **Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters.** *Evolution Int J Org Evolution* 2001, **55(3):**635-640.
17. Corander J, Waldmann P, Marttinen P, Sillanpaa MJ: **BAPS 2: enhanced possibilities for the analysis of genetic population structure.** *Bioinformatics* 2004, **20(15):**2363-2369.
18. Corander J, Waldmann P, Sillanpaa MJ: **Bayesian analysis of genetic differentiation between populations.** *Genetics* 2003, **163(1):**367-374.
19. Dawson KJ, Belkhir K: **A Bayesian approach to the identification of panmictic populations and the assignment of individuals.** *Genet Res* 2001, **78(1):**59-77.
20. Excoffier L, Estoup A, Cornuet JM: **Bayesian analysis of an admixture model with mutations and arbitrarily linked markers.** *Genetics* 2005, **169(3):**1727-1738.
21. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data:linked loci and correlated allele frequencies.** *Genetics* 2003, **164(4):**1567-1587.
22. Fu R, Dey DK, Holsinger KE: **Bayesian models for the analysis of genetic structure when populations are correlated.** *Bioinformatics* 2005, **21(8):**1516-1529.
23. Guillot G, Estoup A, Mortier F, Cosson JF: **A spatial statistical model for landscape genetics.** *Genetics* 2005, **170(3):**1261-1280.
24. Guillot G, Mortier F, Estoup A: **Geneland: A computer package for landscape genetics.** *Molecular Ecology Notes* 2005, **5(3):**708-711.
25. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM: **Control of confounding of genetic associations in stratified populations.** *Am J Hum Genet* 2003, **72(6):**1492-1504.
26. Holsinger KE, Wallace LE: **Bayesian approaches for the analysis of population genetic structure: an example from Platanthera leucophaea (Orchidaceae).** *Mol Ecol* 2004, **13(4):**887-894.
27. Liu N, Wu B, Zhao H: **Inference of population structure using mixture model.** *Technical report* 2005 [http://bioinformatics.med.yale.edu/psmix].
28. Purcell S, Sham P: **Properties of structured association approaches to detecting population stratification.** *Hum Hered* 2004, **58(2):**93-107.
29. 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(2):**466-477.
30. Tang H, Peng J, Wang P, Risch NJ: **Estimation of individual admixture: analytical and study design considerations.** *Genet Epidemiol* 2005, **28(4**289-301 [http://www.fhcrc.org/science/labs/tang].
31. Wang J: **Maximum-likelihood estimation of admixture proportions from genetic data.** *Genetics* 2003, **164(2):**747-765.
32. Manel S, Gaggiotti OE, Waples RS: **Assignment methods: matching biological questions with appropriate techniques.** *TRENDS in Ecology and Evolution* 2005, **20(3):**136-142.
33. Coulon A, Guillot G, Cosson J-F, Angibault JMA, Aulagnier S, Cargnelutti B, Galan M, Hewison AJM: **Genetics structure is influenced by lansdcape features. Empirical evidence from a roe deer population.** *Molecular Ecology Notes* in press.
34. Banks MA, Eichert W: **WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data.** *J Hered* 2000, **91(1):**87-89.
35. 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(4):**1989-2000.
36. Rannala B, Mountain JL: **Detecting immigration by using multilocus genotypes.** *Proc Natl Acad Sci USA* 1997, **94(17):**9197-9201.
37. McKeigue PM, Carpenter JR, Parra EJ, Shriver MD: **Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations.** *Ann Hum Genet* 2000, **64(Pt 2):**171-186.
38. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM: **Design and analysis of admixture mapping studies.** *AmJ Hum Genet* 2004, **74(5):**965-978.
39. Li SL, Yamamoto T, Yoshimoto T, Uchihi R, Mizutani M, Kurimoto Y, Tokunaga K, Jin F, Katsumata Y, Saitou N: **Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci.** *Hum Genet* 2006, **118(6):**695-707.
40. Kuroda Y, Kaga A, Tomooka N, Vaughan DA: **Population genetic structure of Japanese wild soybean (Glycine soja) based on microsatellite variation.** *Mol Ecol* 2006, **15(4):**959-974.
41. Manel S, Bellemain E, Swenson JE, Francois O: **Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited.** *Mol Ecol* 2004, **13(5):**1327-1331.
42. Pearse DE, Arndt AD, Valenzuela N, Miller BA, Cantarelli V, Sites JW Jr: **Estimating population structure under nonequilibrium conditions in a conservation context: continent-wide population genetics of the giant Amazon river turtle, Podocnemis expansa (Chelonia; Podocnemididae).** *Mol Ecol* 2006, **15(4):**985-1006.
43. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society Series B* 1977, **34:**1-38.
44. Yang BZ, Zhao H, Kranzler HR, Gelernter J: **Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE.** *Genet Epidemiol* 2005, **28(4):**302-312.
45. Pritchard JK, Donnelly P: **Case-control studies ofassociation in structured or admixed populations.** *Theor Popul Biol* 2001, **60(3):**227-237.
46. Turakulov R, Easteal S: **Number of SNPS loci needed to detect population structure.** *Hum Hered* 2003, **55(1):**37-45.
47. 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(3):**650-659.
48. Akaike H: **A new look at the statistical identification model.** *IEEE Trans Automatic Control* 1974, **19:**716-723.
49. Zhu X, Zhang S, Zhao H, Cooper RS: **Association mapping, using a mixture model for complex traits.** *Genet Epidemiol* 2002, **23(2):**181-196.
50. Chen H, Chen J, Kalbfleisch JD: **A modied likelihood ratio test for homogeneity in finite mixture models.** *Journal of Royal Statistical Society B* 2001, **63:**19-29.

51. Chen H, Chen J, Kalbfleisch JD: **Testing for a finite mixture model with two components.** *Journal of Royal Statistical Society B* 2004, **66:**95-115.
52. Stephens M: **Dealing with label-switching in mixture models.** *Journal of Royal Statistical Society B* 2000, **62:**795-809.
53. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68(4):**978-989.