Molecular Ecology (2006) **15**, 2833–2843 doi: 10.1111/j.1365-294X.2006.02994.x

Bayesian identification of admixture events using multilocus molecular markers

JUKKA CORANDER and PEKKA MARTTINEN
Department of Mathematics and Statistics, PO Box 68, Fin-00014 University of Helsinki, Finland

Abstract

Bayesian statistical methods for the estimation of hidden genetic structure of populations have gained considerable popularity in the recent years. Utilizing molecular marker data, Bayesian mixture models attempt to identify a hidden population structure by clustering individuals into genetically divergent groups, whereas admixture models target at separating the ancestral sources of the alleles observed in different individuals. We discuss the difficulties involved in the simultaneous estimation of the number of ancestral populations and the levels of admixture in studied individuals' genomes. To resolve this issue, we introduce a computationally efficient method for the identification of admixture events in the population history. Our approach is illustrated by analyses of several challenging real and simulated data sets. The software (BAPS), implementing the methods introduced here, is freely available at http://www.rni.helsinki.fi/~jic/bapspage.html.

Keywords: admixture, Bayesian modelling, genetic mixture, molecular markers, stochastic optimization

Received 16 November 2005; revision received 10 February 2006; accepted 10 April 2006

Introduction

Molecular markers have shown during the past two decades their value and versatility in the detection of the hidden genetic structures of populations. More recently, the focus in such studies has shifted from an exclusive use of traditional genetic distance-based methods increasingly also to utilization of Bayesian model-based approaches (Pritchard *et al*. 2000; Dawson & Belkhir 2001; Corander *et al*. 2003, 2004, 2006; Falush *et al*. 2003). For recent reviews of advances in Bayesian demographic modelling, see Beaumont (2004), and Beaumont & Rannala (2004). However, it has also become clear that data sets representing complex population structures are challenging for any of the Bayesian methods listed above, due to the computational burden imposed by simulation-based inference. In particular, inferences about the number of underlying source populations may be unstable either due to computational difficulties or problems in the specification of the statistical model itself.

There are numerous practical challenges concerning inference about population histories when using molecular methods: (i) the number of available molecular markers may be relatively small, (ii) nonmolecular knowledge about the demographic history for validation of results may be scarce, (iii) samples may represent a wide geographic area comprising many potential source populations, and finally (iv) some statistical methods may require a prohibitive amount of computational resources for the analysis. All these issues are relevant for applications in molecular ecology. Given the challenges, it is necessary to understand the limitations of marker-based demographic inference. Identification of admixture events in the presence of an unknown number of putative ancestral sources in a genetically structured population seems to be among the least tractable issues. For instance, there is a range of examples in the literature involving from only a couple of microsatellite loci (e.g. Heuertz *et al*. 2004) up to hundreds of loci (e.g. Rosenberg *et al*. 2002; Bamshad *et al*. 2003; Rosenberg *et al*. 2005), where formal inference about the number of underlying ancestral populations based on the STRUCTURE software (Pritchard *et al*. 2000; Falush *et al*. 2003) has not yielded biologically sensible results. A common feature for these applications is a tendency to overestimate the number of underlying ancestral sources. Similarly, in a simulation study comparing various clustering software for population genetic analyses, Latch *et al*. (2006) observed in BAPS software (Corander *et al*. 2003, 2004, 2006) a tendency to overestimate the number of

Correspondence: Jukka Corander, Fax: +358 919151400; E-mail: jukka.corander@helsinki.fi

underlying populations when the genetic structure was weak and the analyses were performed at an individual level (instead of sample population level). However, the additional spurious clusters that were obtained typically contained only a very small number of individuals, and could thus not be regarded as relevant estimates of panmictic source populations. Overall, Latch *et al*. (2006) concluded that the statistical power of the most recent BAPS version (Corander *et al*. 2006) was comparable to that of STRUCTURE in detecting genetically differentiated groups in moderately challenging evolutionary scenarios. However, this version of BAPS requires only a fraction of the computational resources necessitated by the Markov chain Monte Carlo (MCMC)-based inference in STRUCTURE. In an average genetic mixture analysis reported by Latch *et al*. (2006), the results were obtained 300–400 times faster with BAPS as compared to STRUCTURE. Therefore, the differences between the two methods become accentuated when the size and the complexity of the investigated data increase.

Here we introduce a numerically viable strategy for a reliable identification of admixture events in the ancestry of sampled individuals using multilocus molecular markers. The need for this strategy becomes apparent when considering complex molecular information for which the earlier approaches to estimation are expected to be neither reliable nor practically feasible. We discuss the difficulties involved in determining simultaneously admixture events and the number of ancestral populations represented in a sample, and suggest that inference about the latter should be established prior to the estimation of admixture. The traditionally used Bayesian formulation makes it tricky to specify a flexible prior distribution for admixture estimation, which would allow both for a wide variety of demographic scenarios and mating systems, and also ensure that the likelihood does not capture random patterns in the data as evidence for admixture events. These goals are achieved here by combining a discrete parameterization of admixture proportions in genomes with a simulation framework that yields a clear biological interpretation of the estimation results and can be used to assess the statistical significance of putative admixture events. It is also shown how a typical geographical sampling scheme can be utilized to strengthen the inferences from weakly informative molecular data. To illustrate our framework, we present analyses of several challenging real and simulated data sets.

Materials and methods

Statistical methods

Earlier works on BAPS have demonstrated the versatility of the Bayesian stochastic partition approach to genetic mixture modelling. In particular, compared to the MCMC-based estimation used in Corander *et al*. (2003, 2004), the stochastic optimization algorithm introduced in Corander *et al*. (2006) improves significantly the applicability of the mixture model to challenging data sets. Corander *et al*. (2006) also introduced a genetic mixture model which allows partial or complete baseline data from putative source populations to be coherently incorporated into the prior specification. In the Appendix, we provide technical details concerning the weak identifiability of an admixture model, where the number of ancestral populations is an unknown parameter. This suggests that the admixture inferences may be sensitive to the choice of a prior distribution and that default options may yield spurious results unless a biologically meaningful number of ancestral populations is used. Given these apparent inferential problems, we adopt a sequential modelling strategy where the number of genetically differentiated sources contributing to a data set is inferred first using a mixture model. Thereafter, given such an estimate, admixture events can be learned on a more stable basis using a Monte Carlo simulation-based algorithm also described in the Appendix.

In the stochastic partition model for a population consisting of k panmictic parts, a putative population structure is represented by a partition $S = (s_1, \ldots, s_k)$, which allocates n sampled individuals into k nonempty clusters. The prior distribution $p(S)$ of the structure parameter can be defined in various ways, depending on the availability of biologically relevant nonmolecular information. First, let $1 < K \leq n$ be an integer specifying an upper limit for the number of panmictic parts thought to be feasible for the investigated population. Then, if no further information is imposed on the prior, a default uninformative choice yields the probability according to:

$$p(S = (s_1, \ldots, s_k)) = \begin{cases} c, & \text{if } k \leq K \\ 0, & \text{otherwise} \end{cases}, \qquad \text{(eqn 1)}$$

which corresponds to the uniform distribution over the partitions with at most K clusters. However, when the molecular information is weak, e.g. the number of available marker loci is small or the markers have low levels of polymorphism, inferences can be strengthened by utilizing the sample design information in the prior specification as in Corander *et al*. (2003). A commonly used sampling strategy is to collect individuals from a number of geographically limited areas, yielding local sample populations. This enables the calculation of the level of genetic differentiation, e.g. using F_{ST} measures. On the other hand, the same information can be used for a further restriction of the prior (1). In fact, when the number of marker loci is extremely small, no sensible inferences can be expected from the mixture clustering without such a restriction. Let $I(S)$ be an indicator function of the compatibility between a sampling design and a putative structure S, i.e. $I(S)$ equals one when all individuals of any local sample population are allocated in

the same arbitrary cluster of S, and zero otherwise. Also, let $I(k \leq K)$ be the indicator function of the number of clusters k not exceeding the threshold value K. A prior respecting the sample design can then be defined as:

$$p(S = (s_1, \ldots, s_k)) = c_0 I(S)\, I(k \leq K), \qquad \text{(eqn 2)}$$

which corresponds to the uniform distribution over the partitions with at most K clusters and where no local sample populations have been divided into several clusters.

Assume that the sampled individuals are genotyped at N_L molecular marker loci, such that the number of distinct alleles at locus j equals $N_{A(j)}$, $j = 1, \ldots, N_L$. A wide variety of marker types, such as microsatellites, amplified fragment length polymorphisms (AFLPs) and single-nucleotide polymorphisms (SNPs), can be employed in our framework equivalently for haploid, diploid or tetraploid individuals. It should be noticed, however, that for dominant markers (such as AFLPs) the inferences are based on modelling the underlying population genotype frequencies instead of the allele frequencies considered for the co-dominant markers. To obtain a tractable Bayesian clustering model, it is assumed that the marker loci are unlinked and that the source populations contributing to the observed sample are in Hardy–Weinberg equilibrium (HWE). These assumptions lead to the posterior distribution over the space of putative clustering solutions:

$$p(S \mid data) = p(data \mid S)p(S) / \sum_{s \in \Theta} p(data \mid S)p(S), \qquad \text{(eqn 3)}$$

where *data* refers to the observed marker genotypes and $p(data \mid S)$ is the *marginal likelihood*, also called the *prior predictive distribution* of the observed data (see Bernardo & Smith 1994). The sum in (3) is over the space Θ of all partitions; however, only the partitions with positive prior probabilities will contribute to the sum. Explicit formulae for $p(data \mid S)$ have earlier been given in Corander *et al*. (2003, 2004, 2006), based on a Multinomial-Dirichlet model for the observed genotype frequencies under HWE and nonlinkage of the markers. Notice that the clusters emerging under the stochastic partition model are considered exchangeable and remain thus completely unlabelled in the statistical model.

It is important to acknowledge the intrinsic difference between the marginal likelihood and likelihood in the regular statistical sense. The former refers to the probabilistic quantification of the information contained in observed data provided by a specific model structure, when the uncertainty about the model parameters has been taken into account. Likelihood, in turn, is usually interpreted as the conditional distribution of the observed data given any fixed parameter configuration. Whereas the maximized likelihood is a nondecreasing function of the degree of model complexity (e.g. increase of k), the marginal likelihood obeys the generic scientific Occam's razor principle according to which simpler theories are to be preferred if they predict empirical data better or equally well as more complex ones (see, e.g. Bernardo & Smith 1994). In the current genetic mixture modelling context this means that the marginal likelihood $p(data \mid S)$ may decrease considerably when k is increased, if the data does not contain decisive support for differences between the allele (or genotype) frequencies of putative underlying populations.

The stochastic optimization algorithm of Corander *et al*. (2006) targets to identify the mode of the posterior distribution over genetic mixture models with a varying k. As the posterior distribution (3) is proportional to the marginal likelihood $p(data \mid S)$ which can be expressed analytically, the maximization procedure corresponds to a search in the space of clustering solutions subject to the prior constraints. Let a 'sampling unit' denote either an individual in the sampled data or an a priori given group of individuals, corresponding to the analyses performed under the prior (1) and (2), respectively. For such sampling units, the following search operators are used to improve the estimate of the genetic structure:

1 Move sampling units from one cluster to another in a stochastic order.
2 Join clusters of sampling units.
3 Split clusters using the Kullback-Leibler (KL) divergence between sampling units (for definition of KL divergence, see Corander *et al*. 2003).
4 Re-allocate several sampling units from a cluster in a random order.

Although these operators are partially similar to those employed in the MCMC-based approaches of Corander *et al*. (2003, 2004), they are not embedded into a Markov chain, which makes their use computationally much more efficient.

Given the posterior mode estimate of the genetic mixture in terms of S, we proceed with the identification of admixture events using an algorithm described in the Appendix. Notice that after the identification of the genetically differentiated parts of a population based on the mixture model, we allow any individual putatively to have an admixed ancestry from any of the source populations. This holds true even if the genetic mixture was inferred using the prior restriction with respect to the local sample populations. Also, if the genetic mixture analysis indicates several roughly equally plausible candidates of the underlying structure, the admixture analysis can be performed separately for each of them. For instance, if the estimated posterior probabilities for $k = 5$ and $k = 6$ are both reasonably high for some data set, the admixture analysis can be done both under the estimated genetic mixture with 5 clusters as well as under that with 6 clusters.

A summary of the features of the admixture estimation is as follows. For any particular individual in the data, let $\boldsymbol{q}$ denote a vector of length k, where the element q_i represents the proportion of the genome inherited from an ancestral source corresponding to the ith cluster in the mixture model. Here we consider solely a discrete version of the admixture coefficients where $q_i \in [0, 0.01, \ldots, 0.99, 1]$, for all $i = 1, \ldots, k$, such that $\Sigma q_i = 1$. This discretization simplifies the inference considerably, and we think that accuracy beyond the level of 1% is not attainable in most applications and is also not important from the perspective of practical interpretation. It is worth noticing that in any admixture model each observed allele represents $[100/(2N_L)]$ percentage of the genome in the likelihood (assuming no missing alleles and a diploid species). For instance, with 10 observed loci for a diploid species, it is not possible even in theory to infer admixture beyond 5% accuracy. In particular, we suggest that caution is needed in quantitative comparisons of admixture coefficients when the number of loci available is only small or moderate.

Given the uniform prior distribution for all $\boldsymbol{q}$-vectors over their finite support, we can numerically maximize the posterior of $\boldsymbol{q}$ for any fixed value of the allele (or genotype) frequencies of the underlying population mixture. By generating realizations from the posterior distribution of the allele frequencies and averaging the admixture estimates over these, we obtain a final estimate $\hat{\boldsymbol{q}}$ of $\boldsymbol{q}$ for each individual, such that the uncertainty about the underlying ancestral allele frequencies has been taken into account.

To assess the significance of the estimated admixture coefficients ($\hat{\boldsymbol{q}}$), we utilize a simulation framework that has a clear interpretation and enables a joint treatment of all individuals simultaneously. Biologically, it is of importance to determine whether the admixture coefficients deviating from zero represent genuine contributions from the corresponding ancestral sources, or whether they, e.g. simply reflect the uncertainty about the allele frequencies in two particularly similar source populations. We assess this issue by simulating multilocus genotypes for, say m, individuals from each source population in the genetic mixture model where the allele frequencies match with the posterior expectation. For each of the simulated individuals, the admixture coefficients are then estimated by the same procedure which was used for the individuals in the observed data. As a consequence, these estimates represent realizations from the distribution corresponding to the null hypothesis of no admixture events. Let $\hat{q}_i$ be the estimated admixture coefficient corresponding to the source where the individual was allocated in the mixture estimation. A P value of obtaining an admixture coefficient less than or equal to $\hat{q}_i$ under the null hypothesis can then be calculated as the proportion of the m realizations for which the condition remains true. This proportion is directly associated to the chance of obtaining at least comparable evidence for admixture under random sampling of non-admixed individuals from the source populations. It is important to notice that the reference individuals are generated from the posterior of the allele frequencies of each inferred ancestral source. Potential differences in the sample sizes from these sources are taken into account since the posterior variability increases with decreasing sample size.

For many population genetic studies, it is essential to ensure that the demographic conclusions are not biased by the presence of individuals representing source populations that are not covered by the sampling design to a sufficient extent. It is clear that a reliable estimation of the ancestral allele frequencies is not feasible by any method for populations which are only represented by a limited number of sampled individuals. Therefore, we have included in our method the possibility of excluding from the admixture inference very small outlier clusters identified in the genetic mixture analysis.

Real data sets

To test our method in a challenging real biological scenario, we performed a re-analysis of the human data set analysed in Rosenberg *et al*. (2002). The data consist of 1056 individuals sampled from 52 different populations around the world. In total, 377 di-, tri- and tetranucleotide microsatellite loci spread over 22 autosomal chromosomes were available. The proportion of missing genotypes is 3.8% and they are fairly uniformly distributed across the loci. Other analyses of human demographic history of a similar magnitude are reported in Bamshad *et al*. (2003) and Rosenberg *et al*. (2005).

In the study by Rosenberg *et al*. (2002), the structure of human populations was investigated by using the STRUCTURE software. The results showed five well-defined groups which seemed to correspond to five major geographic regions, excluding an additional outlier, the Kalash population. However, it was also reported in their supplementary material that the estimation algorithm started to converge to different solutions in separate runs when the number of ancestral sources was specified to be higher than six. Furthermore, the approximate posterior inference about the number of ancestral sources indicated towards a much higher value than seemed plausible given the knowledge about human populations. Therefore, in their global analysis, the number of sources chosen in the displayed main results was not based on formal inference, but on reasoning on the basis of external knowledge.

In all analyses reported here, 100 realizations from the posterior of the allele frequencies were used in the expectation step of our estimation algorithm. Furthermore, to assess the significance of the admixture estimates, 200 individuals were generated from each identified ancestral source to provide an approximation to the distribution of the estimates under the hypothesis of no admixture.

Table 1 Pairwise F_{ST} values between the groups used for simulating individuals in the test scenarios (computed using GDA, Lewis & Zaykin 2002)

	Eurasia	Surui	America	Oceania	Karitiana	East Asia
Surui	0.163					
America	0.063	0.122				
Oceania	0.053	0.211	0.093			
Karitiana	0.128	0.221	0.090	0.176		
East Asia	0.030	0.155	0.054	0.047	0.123	
Africa	0.035	0.190	0.091	0.068	0.154	0.054

Simulated data sets

To establish the validity of our method, we conducted a large-scale simulation study both utilizing the real human molecular data, and also synthetic populations to provide a biologically relevant foundation. First, we used the subpopulations of the human data that were found genetically diverged in the real data analysis. These subpopulations correspond to those identified also by Corander *et al.* (2004) in a sample population level analysis of genetic mixture. F_{ST} values between the subpopulations were computed using GDA (Lewis & Zaykin 2002) and these are presented in Table 1. Posterior means of the allele frequencies of the subpopulations were used as a basis for simulating individuals from these panmictic units under the assumptions of Hardy–Weinberg equilibrium and no linkage among the loci. Second, to assess the statistical properties of our method for more limited marker sets, we generated several synthetic data sets using the EASYPOP software (Balloux 2001).

In the first human simulation scenario, a total of 55 individuals with genotypes over all 377 loci were repeatedly generated for five pairs of subpopulations corresponding to a range of different genetic distances. For both subpopulations in a specific pair, 25 nonadmixed individuals were first simulated. In addition, we simulated three children such that they had one parent in each population, and two grandchildren such that they had three grandparents in one and one grandparent in the other population. The same scenario was repeated for all considered subpopulation pairs. Two further variants of this simulation framework were also investigated by letting the number of nonadmixed individuals decrease ($n = 20$) or increase ($n = 50$).

In the final human simulation scenario, we created a considerably more challenging population setup. In total, 660 individuals were generated from five different subpopulations, such that 70 of these had an admixed background. This setup was repeated using 377, 200, and 100 marker loci, such that for the latter two cases, the loci actually used in the analysis were randomly chosen from the original ones. The true underlying population structure is shown in Fig. 2.

Using the EASYPOP software we generated four populations (A–D) for which less informative marker data were available. A sample of 30 individuals was taken from each of the populations, and in addition, 10 individuals each admixed within either the population pair (A, B) or (C, D) were generated. All these 130 individuals were analysed twice by performing the genetic mixture analysis both on individual and sample group levels. Biologically relevant sample groups were created by dividing the samples from each population randomly into five groups, each consisting of six individuals. Each admixed individual was randomly allocated to a sample group consisting of individuals from either of the two ancestral populations. The same underlying population setup was used for different levels of genetic differentiation and for varying numbers of marker loci (10–20 loci).

Results

Real data sets

For the original human data, the posterior distribution for the number of ancestral sources was completely concentrated on $k = 7$, which is equivalent to the results of a sample population-based analysis reported in Corander *et al*. (2004). In Fig. 1, the estimated admixture proportions are presented. We note that, in general, Fig. 1 is in a close agreement with the global-level analysis of Rosenberg *et al*. (2002). For instance, both analyses suggest that:

1 The clearest dual ancestry for a sampled population is found for Hazara and Uygur, which are of Eurasian–East Asian origin.
2 The proportion of European ancestry among American populations is largest for the Mayas.
3 A considerable proportion of the North African Mozabite population has an African origin, while the main component is European.
4 Biaka Pygmy, Bedouin and Japanese samples each contain a single individual, whose origin clearly deviates from those of the other individuals in the same sample.

The largest differences observed between the two results concern the status of the Kalash population as an outlier and the separation of the American populations. In Rosenberg *et al*. (2002) the American samples were in the global analysis mainly represented by a single ancestral source, whereas they are clearly split into three groups in Fig. 1. The results were more similar to ours when they performed the analysis using only the sampling populations of American origin. It is useful to consider these differences in the light of the analyses reported by Rosenberg *et al*. (2005), where the same individuals were investigated using an extended marker set (a total of 993 loci). These additional

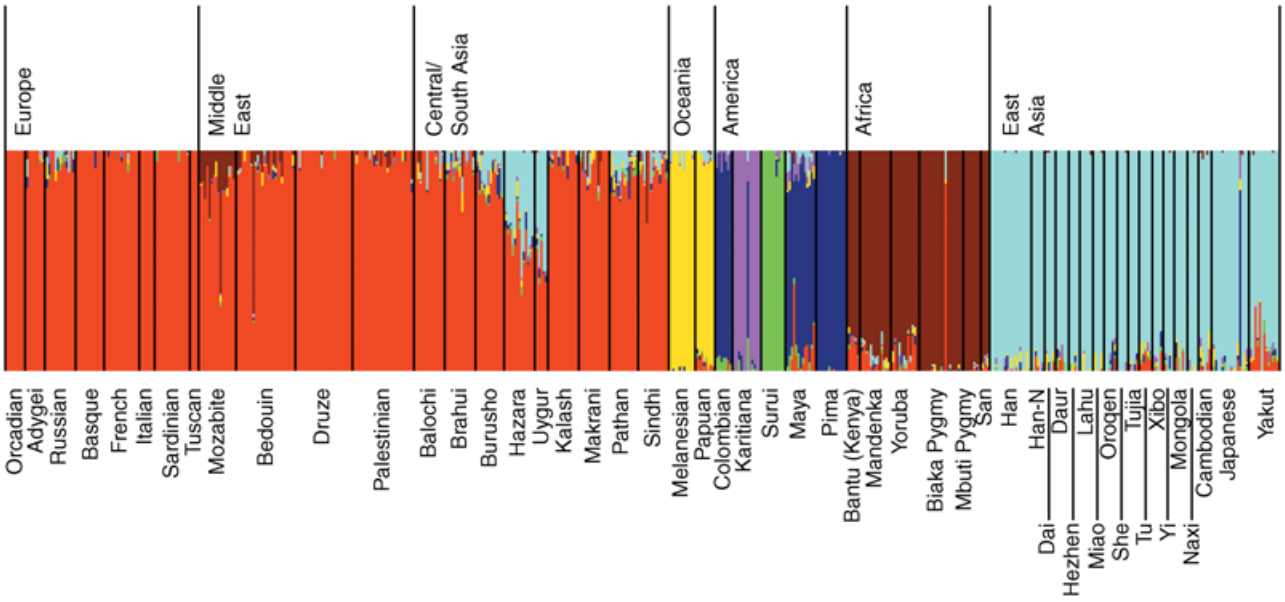

Fig. 1 Estimated admixture coefficients for the human data set from Rosenberg *et al*. (2002). Each column (or vertical line) corresponds to one individual. Ancestral populations are represented by different colours. Each column is coloured with different colours in proportions corresponding to estimated admixture coefficients of the corresponding individual. The sampling populations are separated by black vertical lines and major continental regions by the vertical lines above the coloured area.

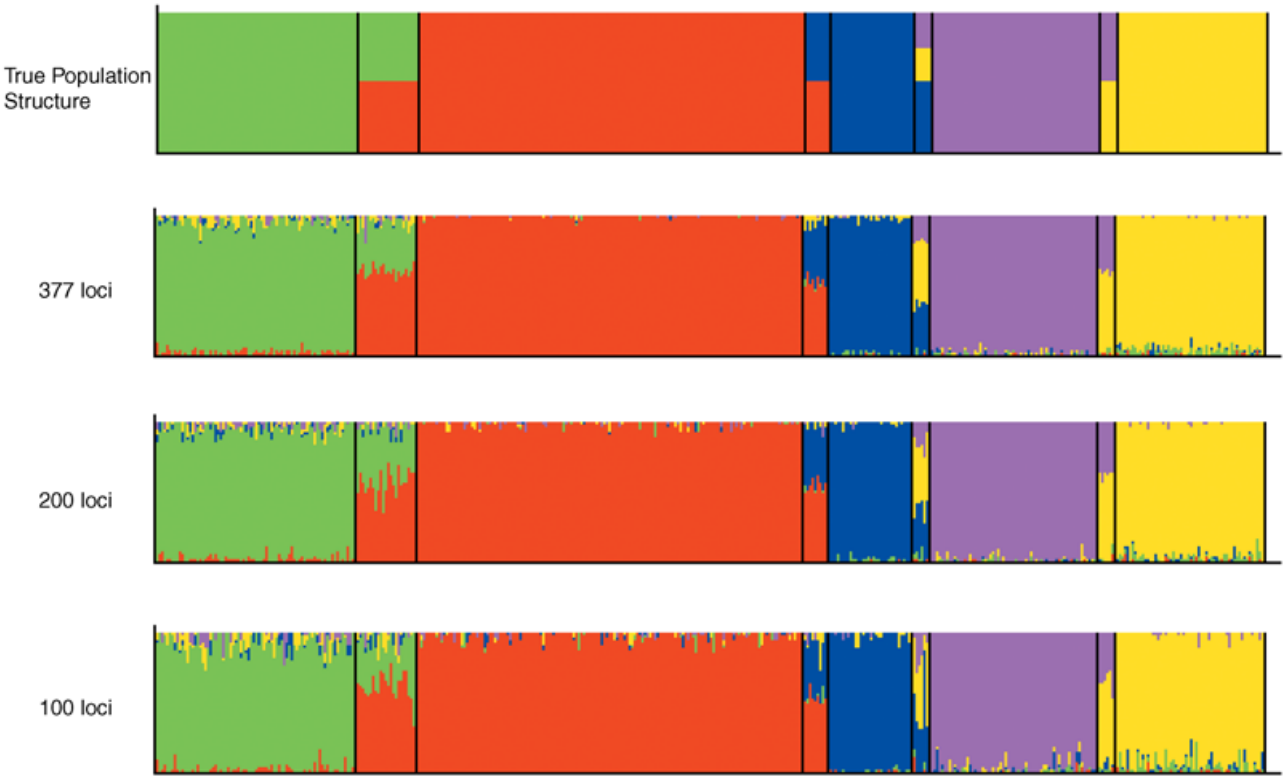

Fig. 2 Results from a simulation scenario with 660 individuals. The same individuals were analysed using different numbers of loci. When the number of loci was less than 377, the loci used in the analysis were randomly chosen from the original ones. The individuals were simulated from five different populations: Eurasia (green), Africa (red), America (blue), Oceania (purple), East Asia (yellow).

data show that the earlier reported separation of the Kalash population was presumably an artefact caused by the numerical instability of the inference, and that the Americas region should be considered heterogeneous even when the molecular variation is investigated at a global level.

When we examined the P values computed for each individual, we found out that 156 out of the total 1056 individuals had values less than 0.05, indicating at least moderate evidence for admixed background. The individuals in the sample populations Hazara and Uygur were all considered as admixed. In fact, apart from one individual in Hazara, P values for all individuals in both of these populations were very close to zero. Other sampling populations for which a majority of individuals had significant admixture coefficients were Mozabite, Burusho, Pathan, Maya, and Bantu-Kenya. The Colombian population can be taken as an example of a case in which examining merely the admixture coefficients might imply an admixed background for the individuals, but where the P values are not supporting such a conclusion. The estimated proportion of admixed background is around 10% for the Colombian individuals. However, the P values for the Colombian individuals range between 0.115 and 0.995 indicating that such an amount of admixture could be regarded as a result of random variation in the genotype patterns. This example illustrates that admixture coefficients can attain nonzero values also for genetically relatively similar alternative ancestral populations which are still differentiated from the identified primary ancestral population of an individual. This is due to the shape of the admixture likelihood, which may under such circumstances yield for a random set of loci a better score for classification of an allele to an alternative ancestral population than to the primary source population of the particular individual. The P values thus enable the separation of spurious evidence for admixture from more conclusive patterns in the marker data, which is difficult to achieve in a strictly Bayesian analysis using information solely from the posterior distribution of the coefficients.

Simulated data sets

For each simulated human data set, our sequential inference method was able to identify the correct number of ancestral populations. The results for the first simulation scenario are presented in Table 2. It is seen that regardless of the genetic distance of the two source populations, the identification of admixed individuals using the 0.05 level of the P value was perfect. However, the estimated values of the admixture coefficients of the admixed individuals had in some cases a considerable deviation from the correct value, and this deviation seemed to increase slightly when the two populations were closer to each other. When the number of nonadmixed individuals from each population was increased ($n = 50$) or decreased ($n = 20$), the results were generally in agreement with those given in Table 2. However, for the larger samples the estimates of the admixture coefficients were closer to the underlying true values. Also, when only 20 nonadmixed individuals were simulated from the subpopulations associated with the smallest genetic distance (Eurasia and Africa), a single cluster emerged in the first step of the analysis, thus leaving the admixture events undetected.

Table 2 Results of admixture estimation for data simulated from two distinct human populations

Population 1	Population 2	F_{ST}	I	II	III	IV	V
Karitiana	Surui	0.221	100	100	< 0.001	0.10	0.06
America	Surui	0.122	100	100	0.005	0.11	0.07
Oceania	America	0.093	100	100	0.013	0.13	0.09
Oceania	Eurasia	0.053	100	100	0.034	0.20	0.08
Eurasia	Africa	0.035	100	100	0.030	0.16	0.09

Three types of individuals were simulated; 25 nonadmixed individuals from each of the two populations, three individuals having one parent in each of the two populations, and two individuals having three grandparents in one and one grandparent in the other population. The simulation was performed five times using different pairs of ancestral populations. The columns presenting the results are as follows: I, percentage of correct identification of nonadmixed individuals, using 0.05 level of P value; II, percentage of correct identification of admixed individuals using 0.05 level of P value; III, average absolute error in the admixture coefficients of nonadmixed individuals; IV, average deviation of admixture coefficients from 0.5 for the second type of simulated individuals; and V, average deviation of admixture coefficients from 0.25 (or 0.75) for the third type of simulated individuals.

In the simulation framework with five underlying ancestral sources and 377 loci (see Fig. 2), all the admixed individuals were correctly identified when using level 0.05 as threshold for the P value. However, two of the nonadmixed individuals were also incorrectly detected as having significant admixture coefficients. When 200 loci were used, the identification of admixed individuals was still complete, but four nonadmixed individuals were now recognized as having admixed background. Finally, with 100 loci the accuracy of identifying admixed individuals decreased a little. Still only one individual having an admixed background was falsely assigned nonsignificant admixture coefficients, but the number of nonadmixed individuals having significant admixture coefficients was 11.

The simulations utilizing EASYPOP illustrate the apparent challenge of inferring admixture events from a limited marker set for weakly differentiated populations (Table 3). When the markers are very polymorphic under such

Table 3 Results of admixture estimation for data simulated with EASYPOP

			Individual level mixture analysis			Group level mixture analysis		
F_{ST}	#Loc	#All	#Outliers	Admix	Nonadmix	#Outliers	Admix	Nonadmix
0.120	20	5	0	7/10	117/120	0	9/10	119/120
0.097	20	15	6	4/4	118/120	0	10/10	119/120
0.057	10	15	17	1/6	107/107	0	6/10	117/120

Four different populations A–D were simulated under three distinct configurations with varying level of differentiation and molecular information. A sample of 30 individuals was generated from each of the four populations, and in addition, 10 individuals each admixed within either the population pair (A, B) or (C, D) were created. All these 130 individuals were analysed twice with our sequential method, by performing the genetic mixture analysis both on individual and sample group levels. The local sample groups were created by dividing the samples from each population randomly into five groups, each consisting thus of six individuals. Each admixed individual was randomly allocated to a sample group consisting of individuals from either of the two ancestral populations. The same underlying population setup was used for three different levels of genetic differentiation and a varying number of marker loci (10–20 loci). The columns refer to: F_{ST}, average pairwise F_{ST} distance between the four simulated populations; #Loc, number of simulated loci in the data; #All, number of possible distinct allelic forms at any locus; #Outliers, number of individuals that were allocated in genetic mixture analysis to outlier clusters consisting of less than five individuals; Admix, x/y, y is the number of admixed individuals that were analysed in admixture stage (not outlier individuals), x is the number of admixed individuals who had significant *P* values for admixture; Nonadmix, x/y, y as in Admix, but for the nonadmixed individuals, y is the number of nonadmixed individuals having nonsignificant *P* values for admixture. The significance level of 5% was used.

circumstances, some individuals can be assigned to small outlier clusters. Nevertheless, all major clusters detected in our simulation studies corresponded to real underlying populations. The default option in the BAPS admixture estimation module is to exclude clusters containing less than five individuals, and hence, in cases where an admixed individual was assigned to an outlier cluster, the corresponding admixture event remained undetected. However, Table 3 illustrates well the advantages of conditioning the inference on relevant local sample populations, as the accuracy of detecting the true underlying population structure and the admixture events increases considerably for the most difficult scenarios. Notice that the actual simulated molecular information was exactly the same for the individual level and sample group level analyses. The local sample populations used here were quite small, and correspond thus to an only sparsely informative geographic sampling design.

Discussion

Experiences from the development of our current approach and applications of previously introduced methods for detecting admixture events show unquestionably that such inference is considerably less tractable than genetic mixture analysis. Nevertheless, when sufficiently informative marker data are combined with an appropriate statistical approach, the accuracy of the inferences can be fairly high. From a theoretical perspective, it is widely accepted that a purely Bayesian statistical approach offers efficient means for a solid characterization of the information contained in empirical data. However, our experiences suggest that it may be too great a challenge to determine widely applicable prior distributions that are expected to provide biologically relevant formal inferences simultaneously about admixture events and the number of ancestral populations. In this respect, the actual numerical applicability of any proposed methodology should not be ignored either. Since the posterior inferences can be dominated by the likelihood in the presence of uninformative priors, we developed the simulation based assessment method to avoid spurious support for admixture events, e.g. when two ancestral populations are only weakly differentiated. All performed simulation studies show that our method is capable of preserving a low false positive rate concerning the admixture events, while being even slightly conservative compared to the nominal significance level.

Although we have focused here on the admixture estimation based on the mode estimate of the genetic structure from a mixture analysis, in the BAPS implementation it is also possible to do inference conditional on either predefined source populations or other clustering solutions associated with high posterior values (these can be inserted as predefined populations). The simulation studies showed the utility of conditioning the genetic mixture inference on local sample populations for weakly differentiated populations and very limited marker sets (as was done in Corander *et al*. 2003). In practice, it is quite feasible to do mixture analyses both at the individual and sample group level (when the latter information exists), since the inference algorithms are fast. This strategy offers the possibility of assessing the relevance of the results from a biological perspective, and may provide further insights about the

possibility of detecting admixture events. For some applications, it may also be fruitful to utilize EASYPOP to simulate population data with the same level of differentiation and marker information as observed in the real data. When such data sets are analysed with BAPS, one can gain information about the expected accuracy of inferences for the real data. EASYPOP produces data sets based on the GENEPOP (Raymond & Rousset 1995) data format, which can be straightforwardly used for BAPS analyses.

The central rationale behind the separation of mixture and admixture analyses is the importance of identifying the number of diverged groups hidden in the data. For instance, outlier groups containing only a single or very few individuals can then be excluded from an admixture analysis, since the data would not allow a reliable estimation of the ancestral allele frequencies originating from such sources. However, it is clear that the power to detect certain types of admixture scenarios is limited for the modelling approach discussed here. For instance, it is unlikely that one could reliably infer situations where one or several population samples consist of only admixed individuals in which the ancestral sources as such are not represented at all. For successful admixture identification, the molecular data should be concisely informative about the events in the population history. Since the estimation of admixture proportions is meaningful only when conditioned on a plausible number of ancestral sources, the importance of using a reliable estimate for that purpose is stressed. It should be noticed that some bias is introduced by the sequential procedure, since when an individual has an admixed background, only a part of the observed alleles should affect the posterior of the allele frequencies in the group to which the individual is assigned. When there is a sample of a reasonable size containing nonadmixed or only moderately admixed individuals available from each considered ancestral source, such biases will in practice be small.

In our current work, we have been able to design computationally attractive algorithms for an investigation of the genetic composition of a sample using multilocus molecular markers. For instance, the whole sequential admixture estimation procedure for the complete human data, which represents quite an extreme evolutionary scenario, took approximately 30 min on a PC with a 2.8GHz Pentium 4 processor. The performance of the algorithms was very stable when the human data was repeatedly analysed using 100 replicate values of the upper bound K, varying it between 10 and 30. The estimation runs yielded almost identical results, and indicated all the differentiated populations presented in Fig. 1. Given the evidence from analyses of both simulated and real data, we expect that the advances introduced here facilitate routine application of the Bayesian approach in investigations of the genetic structure of populations.

Acknowledgements

This work was supported by the Centre of Population Genetic Analyses, University of Oulu, Finland (Academy of Finland, grant no. 53297) and by the research funds of University of Helsinki.

References

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, 92, 301–302.

Bamshad MJ, Wooding S, Watkins WS *et al.* (2003) Human population genetic structure and inference of group membership. *American Journal of Human Genetics*, **72**, 578–589.

Beaumont MA (2004) Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity*, **92**, 365–379.

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251–261.

Bernardo JM, Smith AFM (1994) *Bayesian Theory*. Wiley, Chichester, UK.

Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. *Fishery Bulletin*, in press.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fletcher R (1987) *Practical Methods of Optimization*. Wiley, New York.

Heuertz M, Hausman JF, Hardy OJ, Vendramin GG, Frascaria-Lacoste N, Vekemans X (2004) Nuclear microsatellites reveal contrasting patterns of genetic structure between western and southeastern European populations of the common ash (*Fraxinus excelsior* L.). *Evolution*, **58**, 976–988.

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, DOI: 10.1007/S10 592-005-9058-1.

Lewis PO, Zaykin D (2002) *Genetic Data Analysis: Computer Program for the Analysis of Allelic Data*. Version 1.1. http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php.

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

Raymond M, Rousset F (1995) GENEPOP: population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, **1**, doi: 10.1371/journal.pgen.0010070.

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

The authors are interested in Bayesian Modelling and Molecular Evolution.

Appendix

Here we provide details about the admixture estimation algorithm, and also, the vague identifiability of the number of ancestral sources contributing to a data set under an admixture model is discussed.

Identifiability of the number of ancestral sources contributing to data

Under the assumptions stated in the Materials and Methods section, it is straightforward to extend the stochastic partition representing a genetic mixture in Corander *et al*. (2004, 2006) to a representation of a potentially admixed genetic composition of a sample. In the earlier stochastic partition models corresponding to a genetic mixture, every allele observed from a specific individual assigned to a cluster, say i, is restricted to represent the allelic composition of that source population. Mathematically, for a fixed partition with k clusters, such a representation is similar to the likelihood arising under the latent class mixture model (with k classes) used in STRUCTURE. The partition model may be generalized by allowing the alleles of an individual to be assigned to separate source populations, which leads to a similar likelihood as that arising under the latent class admixture model (again with k classes) used in STRUCTURE.

Using a slightly different definition from that in the Materials and Methods section, let $S = (s_1, \ldots, s_k)$ now represent a partition of all observed alleles, from all n individuals, into k nonempty clusters. Each s_i represents then an ancestral source contributing to the sample in terms of certain alleles over the considered loci. For instance, when one-half of the observed alleles of an individual are allocated to a particular s_i, and the remaining alleles to another ancestral source, the individual is considered to have one parent in each of these source populations. Analogously to Corander *et al*. (2004, 2006) the marginal likelihood of the molecular data, conditional on the partition S, equals then:

$$p(data|S) = \prod_{i=1}^{k}\prod_{j=1}^{N_L}\left[\frac{\Gamma\left(\sum_l \alpha_j\right)}{\Gamma\left(\sum_l (\alpha_j + n_{ijl})\right)}\prod_{l=1}^{N_{A(j)}}\frac{\Gamma(\alpha_j + n_{ijl})}{\Gamma(\alpha_j)}\right], \quad \text{(eqn A1)}$$

where $\Gamma(\cdot)$ is the gamma function, n_{ijl} is the number of copies of allele l ($l = 1, \ldots, N_{A(j)}$) at locus j considered to emerge from ancestral source i, and α_j is a Dirichlet prior hyperparameter. The statistical model underlying the marginal likelihood is a product Multinomial-Dirichlet distribution over the loci and ancestral sources, where the probabilities p_{ijl} represent the unknown allele frequencies of the ancestral populations (these are integrated out in equation A1). Notice that there is no proportionality sign in (A1) unlike in the comparable formula of Corander *et al*. (2003). This is due to derivation directly in terms of a generalization of the deFinetti representation theorem (see, e.g. Bernardo & Smith 1994).

The analytical result given in (A1) enables us to investigate some important properties of the stochastic partition model in the admixture case. In particular, if no alleles are allocated to ancestral population i at locus j, the corresponding term in the marginal likelihood equals unity for any value of the hyperparameter. Thus, the value of (A1) remains the same when the alleles at locus j allocated to the ancestral source i are moved to another arbitrary ancestral source, which does not yet have any allocated alleles *at that particular locus*. It follows that the predictive power of the Bayesian model is not lowered by an increase in the value of k, as the alleles at any locus can always be allocated to new ancestral sources not having observations for the particular locus or any loci. This is in sharp contrast with the behaviour of the stochastic partition model for a genetic mixture, where an increase in k always induces also an increase in the effective number of parameters of the model. Similarly, the value of (A1) remains the same, whenever two ancestral sources exchange at any locus the alleles that are allocated there.

These remarks illustrate the important role of the prior for the admixture configuration of an individual. Since the marginal likelihood is under many configurations invariant with respect to an increase in k, an admixture model with an a priori unknown k can be considered as vaguely identifiable. In particular, when the data are extensive and represent a complex population structure with many genetically diverged subpopulations, it may happen that even a reasonable prior for the admixture configuration of an individual has only a negligible contribution to the final inference. We suggest that this vague identifiability explains the paradoxical behaviour observed, e.g. by Heuertz *et al*. (2004) and Rosenberg *et al*. (2002, 2005) with respect to inference about k in the admixture estimation.

Estimation of admixture coefficients

Since it is difficult to specify reasonable priors for the admixture configuration such that the inferences could be expected to be stable for a wide range of biological data, we have developed an alternative strategy to admixture modelling. This is based on two distinct phases where the inference about k is settled first using the genetic mixture model and estimation algorithm of Corander *et al*. (2006), and then, the admixture configuration is estimated for each individual. In the latter stage, we use Monte Carlo integration combined with a discretized version of standard constrained steepest descent algorithm (see, e.g. Fletcher

1987) which provides a numerically extremely fast method of obtaining the marginal maximum a posteriori estimates of admixture coefficients. The Bayesian method of Corander *et al*. (2006) only requires an upper bound for k to be set as a hyperparameter, whereas the other hyperparameters are chosen using reference priors well-established in the literature. Since an upper bound for the number of genetically diverged sources contributing to a sample is a concrete quantity, it is reasonable to formulate a specific opinion about it.

Given an identified genetic mixture with k components, let p_{ijl} ($i = 1, \ldots, k; j = 1, \ldots, N_L; l = 1, \ldots, N_{A(j)}$) be any fixed values of the allele frequencies over the loci and components. In the Monte Carlo integration step p_{ijl} is represented by a realization from the product Dirichlet posterior distribution of the allele frequencies for each $i = 1, \ldots, k$. To simplify the notation, we consider below a single individual only, since the estimation can be performed analogously for all individuals. The vector $\boldsymbol{q} = (q_1, q_2, \ldots, q_k)$, specifies the proportions of the ancestral origins (i.e. admixture coefficients) for a particular individual, with the restrictions: $0 \le q_i \le 1$, $\Sigma q_i = 1$. Denote the observed alleles at locus j by vector (d_{jz}) where the range of z depends on the number of alleles known, say N_{all}. For instance, for a diploid individual $N_{all} = 2$, when there are no missing data at locus j. Let d denote the joint set of (d_{jz}) over loci. Assuming $\boldsymbol{q}$ known, the probability of observing a particular d_{jz} is defined as:

$$p(d_{jz} \mid \mathbf{q}) = \sum_{i=1}^{k} q_i p_{ij(d_{jz})}. \qquad \text{(eqn A2)}$$

Under the assumptions stated earlier for the mixture model, we obtain the following likelihood for all observations for an individual:

$$p(d \mid \boldsymbol{q}) = \prod_{j=1}^{N_L} \prod_{z=1}^{N_{all}} \left(\sum_{i=1}^{k} q_i p_{ij(d_{jz})} \right). \qquad \text{(eqn A3)}$$

Here we consider solely a discrete version of the admixture coefficients where $q_i \in [0, .01, \ldots, .99, 1]$, for all $i = 1, \ldots, k$, such that $\Sigma q_i = 1$. The prior for all $\mathbf{q}$-vectors is assumed uniform in the finite support, which leads to the conditional posterior:

$$p(\mathbf{q} \mid d) = \frac{P(d \mid \boldsymbol{q})}{\sum_{\boldsymbol{q} \in Q} P(d \mid \boldsymbol{q})}. \qquad \text{(eqn A4)}$$

In the maximization step of our algorithm, we obtain the conditional posterior mode of the admixture coefficients for each individual. The marginal maximum a posteriori estimate of $\boldsymbol{q}$ is then given as the mean over the estimates corresponding to the m realizations of p_{ijl} ($i = 1, \ldots, k; j = 1, \ldots, N_L; l = 1, \ldots, N_{A(j)}$) from the product Dirichlet posterior obtained through the mixture partition model.