The computer program Structure for assigning individuals to populations: easy to use but easier to misuse

Jinliang Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J Wang

Right running head: Population Structure Analysis

Key words: Genetic structure, markers, genetic differentiation, admixture, population clustering

Corresponding author:

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

ABSTRACT

The computer program Structure implements a Bayesian method, based on a population genetics model, to assign individuals to their source populations using genetic marker data. It is widely applied in the fields of ecology, evolutionary biology, human genetics and conservation biology for detecting hidden genetic structures, inferring the most likely number of populations (*K*), assigning individuals to source populations, and estimating admixture and migration rates. Recently, several simulation studies repeatedly concluded that the program yields erroneous inferences when samples from different populations are highly unbalanced in size. Analysing both simulated and empirical datasets, this study confirms that Structure indeed yields poor individual assignments to source populations and gives frequently incorrect estimates of *K* when sampling is unbalanced. However, this poor performance is mainly caused by the adoption of the default ancestry prior, which assumes all source populations contribute equally to the pooled sample of individuals. When the alternative ancestry prior, which allows for unequal representations of the source populations by the sample, is adopted, accurate individual assignments could be obtained even if sampling is highly unbalanced. The alternative prior also improves the inference of *K* by two estimators, albeit the improvement is not as much as that in individual assignments to populations. For the difficult case of many populations and unbalanced sampling, a rarely used parameter combination of the alternative ancestry prior, an initial ALPHA value much smaller than the default and the uncorrelated allele frequency model is required for Structure to yield accurate inferences. I conclude that Structure is easy to use but is easier to misuse because of its complicated genetic model and many parameter (prior) options which may not be obvious to choose, and suggest using multiple plausible models (parameters) and *K* estimators in conducting comparative and exploratory Structure analysis.

Introduction

Pritchard and coworkers (Pritchard *et al*. 2000; Falush *et al*. 2003, 2007; Hubisz *et al*. 2009) developed a Bayesian method to assign individuals with multilocus genotypes into discrete clusters, each corresponding to a Mendelian population characterized by a set of allele frequencies at each locus. The method is based on a population genetics model, and yields parameter estimates and assignment results with well-defined and easy-interpretable biological meanings. The method, implemented in the computer program Structure (Pritchard *et al*. 2000), has been widely applied in the fields of ecology, evolutionary biology, human genetics, and conservation biology. It proves to be highly popular, being cited tens of thousands of times by published scientific papers (Puechmaille 2016). Among other purposes, the method has been extensively used to detect hidden

population structure, to estimate the most likely number of populations contributing to a sample of individuals, to assign individuals to their source populations, to infer admixtures (hybridizations) and migrations between inferred populations (Porras-Hurtado *et al*. 2013).

Several recent simulation studies concluded, however, that Structure does not reliably recover the actual population structure when sampling is uneven among populations (e.g. Kalinowski 2011; Neophytou 2014; Puechmaille 2016). They demonstrated that, when the samples from different source populations are highly unbalanced in sizes, Structure tends to underestimate the number of contributing populations and merge populations represented by small samples. They also showed that the pathological results of Structure are not caused by the lack of marker information or the lack of population differentiation, because their simulations used 1000 highly polymorphic microsatellites to infer the structure of populations with high F_{ST} values of ~0.1 (Kalinowski 2011).

If the problem identified by the simulation studies were true, then Structure would be seriously questioned as an efficient or even appropriate tool because real data rarely have balanced sample sizes. One of the strengths of Structure is its ability to reveal, using purely genetic data, cryptic or hidden population structures that are difficult to detect using visible characters such as sampling locations or phenotypic traits (Pritchard *et al*. 2000). This is in contrast to the traditional population genetic structure analysis approaches such as F_{ST} analysis (e.g. Weir & Cockerham 1984) which rely on predefined populations. A typical example is the mixed stock analysis (Smouse *et al*. 1990), where individuals are sampled from the same location but are contributed by an unknown number of phenotypically indistinguishable but genetically differentiated source populations. In such a situation, we have no idea of the populations represented by the samples, let alone the sizes of the samples from the populations. Indeed, if Structure relied on balanced sample sizes to obtain reliable results, then its usefulness would be greatly compromised in practice.

Being a Bayesian method, Structure bases its inferences on various priors as well as genotype data (Pritchard *et al*. 2000). One of the priors is used to model individual ancestry distributions among populations (Falush *et al*. 2003). In Structure's admixture model of K assumed populations, the prior ancestry of individual i being from population j (j=1, 2, …, K), $q_j^{(i)}$, follows the Dirichlet distribution $q^{(i)} \sim \mathcal{D}(\alpha_1, \alpha_2, \cdots, \alpha_K)$, where $q^{(i)} = (q_1^{(i)}, q_2^{(i)}, \cdots, q_K^{(i)})$, $0 \leq q_j^{(i)} \leq 1$, and $\sum_{j=1}^{K} q_j^{(i)} = 1$. Structure has two options for this ancestry prior, "Use a Uniform Prior for α" and "Separate α for each Population". The first option is the default of the program, which uses a single α value for all assumed K populations (i.e. $\alpha_1 = \alpha_2 = \cdots = \alpha_K$). It specifies that each

individual has its ancestry originating from each of the assumed *K* populations at an equal prior probability of 1/*K* (i.e. $q_1^{(i)} = q_2^{(i)} = \cdots = q_K^{(i)} = 1/K$ in expectation). The second option assumes distinct α values for the assumed *K* populations, and an individual may have its ancestry originating from the assumed *K* populations at *K* different prior probabilities (proportions). Structure estimates the one α value (option 1) or *K* different α values (option 2), and uses the one or *K* estimated α values in assigning individuals to populations. Obviously, the default option suits for balanced sampling, but becomes increasingly inappropriate with an increasing difference in the sizes of samples from different populations. The alternative prior applies to unbalanced as well as balanced sampling, although it could incur some cost in accuracy when applied to balanced sampling. A close inspection of the simulation studies (Kalinowski 2011; Neophytou 2014; Puechmaille 2016) described above shows that they invariably used the default ancestry prior in analysing the simulated data. In other words, it could be the misuse of the Structure program that led to the conclusion that the program does not reliably recover the actual population structure when sampling is uneven.

This study has six objectives. First, I use simulations to confirm the conclusion of previous simulations that Structure does not perform well with unbalanced sampling when the default ancestry prior is used. Second, I demonstrate that Structure does yield accurate individual assignments to populations in the presence of highly unbalanced sampling when the alternative ancestry prior is used and when the number of populations is not very large. Third, I show that the alternative ancestry prior also improves the inferred number of populations, *K*. Fourth, I show that the alternative ancestry prior has no detectable cost in inference accuracy when sampling is balanced. Fifth, I show that, combined with some further prior parameter and model adjustments, the alternative ancestry prior makes accurate individual assignments to populations when *K* is very large (say, $K > 40$) and sampling is highly unbalanced. In contrast, the default ancestry prior always yields poor results in such a situation. Sixth, I analyse a human dataset comparatively with the two ancestry priors to show the importance of choosing the right prior in practice. In conclusion, I recommend the wide use of the alternative ancestry prior in Structure analysis. I suggest that, when sampling is (or is suspected to be) highly unbalanced, caution must be exercised about the inferred *K*, and about the prior and parameter choices in conducting Structure analysis when *K* is large. I also encourage the use of Pritchard *et al*. (2000) original method for inferring the most likely number of populations in place of, or in additional to, the popular *ΔK* method proposed by Evanno *et al*. (2005).

Methods

Simulations: The power and accuracy of Structure analyses depend on, among other factors, marker informativeness for individual ancestry (Rosenberg *et al*. 2003) or relatedness (Wang 2006), the pattern (e.g. island, stepping-stone, and isolation-by-distance models; hierarchical structures) and extent of genetic differentiation among populations, and the sampling scheme and sampling strength of individuals from the populations. This study focusses on unbalanced sampling, and considers its impact on Structure analysis when it is expected to be powerful: highly informative marker data from highly differentiated populations in the simple island model.

I assumed the simple situation of a number of K discrete populations in Wright's (1931) island migration model. The populations had reached equilibrium among mutation, drift and migration, where drift was the dominating evolutionary force leading to a high equilibrium F_{ST} value. A number of n_i individuals were drawn at random from population i (i=1, 2, …, K), and each sampled individual was genotyped at a number of L microsatellite loci, each having M codominant alleles. In the simulations presented in this study, M was fixed at 10, but other M values did not change the conclusions.

For a given locus l (l =1, 2, …, L), the ancestral allele frequencies, $\boldsymbol{p_{0l}}$ ={p_{0l1}, p_{0l2},…, p_{0lM}}, were drawn from a uniform Dirichlet distribution $\mathcal{D}(1,1,\ldots,1)$. Conditional on $\boldsymbol{p_{0l}}$, the allele frequencies of a population i, $\boldsymbol{p_{il}}$ ={p_{il1}, p_{il2},…, p_{ilM}}, were drawn from the Dirichlet distribution $\mathcal{D}(fp_{0l1}, fp_{0l2}, \ldots, fp_{0lM})$, where $f = \frac{1}{F_{ST}} - 1$ (Nicholson *et al*. 2002; Falush *et al*. 2003) and F_{ST} is the assumed equilibrium genetic differentiation among the K populations. Given $\boldsymbol{p_{il}}$, a diploid genotype at locus l was drawn at random from population i at Hardy-Weinberg equilibrium, by sampling two alleles independently. The multilocus genotype of a sampled individual was obtained by combining single locus genotypes independently, assuming linkage equilibrium. The $\mathbf{n}$={n_1, n_2, …, n_K} multilocus genotypes were then pooled and subjected to Structure analysis.

My simulations considered different numbers of populations (K=3, 6, 12, 24, 48), loci (L=10, 20, 40, 50) and sampled individuals per population (n_i=5, 10, …, 1365), and different differentiation levels (F_{ST}=0.05, 0.1, 0.2) among populations. For each parameter combination, 100 replicate datasets were simulated and analysed by Structure under different prior and parameter settings as detailed below.

Structure analysis: The simulated data were analysed by the program Structure 2.3.4 (Pritchard *et al*. 2000) to infer the number of populations and to assign individuals to the inferred populations. For each Structure analysis, I used the admixture model and the correlated allele frequency model (Falush *et al*. 2003), as were used by the above described simulation studies (Kalinowski 2011;

Neophytou 2014; Puechmaille 2016) and frequently in empirical data analyses. My preliminary simulations showed that Structure becomes increasingly sensitive to the assumed allele frequency models with an increasing *K*. For the case of many populations, therefore, I also used the uncorrelated allele frequency model for comparison with the correlated allele frequency model. I used a burn-in length of 10^4 and 5×10^5 iterations in analysing data simulated with K=3 and $K\geq$6 populations, respectively. In both cases, a run length of 10^4 iterations was run after the burn-in. Increasing either the burn-in (up to 2×10^6) or the run length did not substantially change the results in test datasets. All other parameters in Structure were left as default, except for the prior of individual ancestry, the initial α value, ALPHA, and the allele frequency model.

Each simulated dataset was analysed comparatively using the default or the alternative prior of individual ancestry. As explained in the *Introduction*, the default prior assumes all presumed populations contribute equally to the sampled individuals. With this prior, Structure adopts the Dirichlet distribution $q^{(i)}\sim\mathcal{D}(\alpha,\alpha,\cdots,\alpha)$ for the prior ancestry distribution of an individual *i*, and estimates the single parameter α in this distribution jointly with other parameters. The alternative prior assumes that different populations may contribute variably to the sampled individuals. In other words, the prior proportional (or probability of) ancestry of an individual *i* from population *j* (*j*=1, 2, …, *K*), $q_j^{(i)}$, may vary with *j*. With this prior under an assumed number of *K* populations, Structure uses the Dirichlet distribution $q^{(i)}\sim\mathcal{D}(\alpha_1,\alpha_2,\cdots,\alpha_K)$ for the prior ancestry of an individual *i*, and estimates the *K* parameters α_j (*j*=1, 2, …, *K*) in this distribution jointly with other parameters.

For both the default and alternative priors, Structure program requires an initial value of α, ALPHA, to start a Markov chain Monte Carlo which is used to update estimates of α and other parameters. The default value of ALPHA is 1.0, which was used, except when explicitly stated, in analysing the simulated and empirical datasets. However, I found by simulations that the inferences of both *K* and individual ancestry by Structure are affected by ALPHA, and are increasingly so with an increase in *K*. For the case of many populations, therefore, different ALPHA values (0.03125~1.0) were also used in both the default and the alternative ancestry priors to show this ALPHA effect in general, and to show that both the alternative prior and a much smaller ALPHA value than the default (1.0) are required for Structure to deliver accurate inferences of *K* and individual ancestry in the case of unbalanced sampling from many populations.

For assessing the accuracy of individual ancestry assignments, a number of 20 independent runs were conducted for each dataset, assuming a $\widehat{K}$ value equal to the simulated *K*. The average assignment accuracy (see below) was assessed across runs and across datasets. For assessing the

accuracy of estimated number of populations $\widehat{K}$, 20 runs were carried out for each assumed $\widehat{K}$ value ranging from *K*-2 to *K*+2 for each dataset. The most likely number of populations was inferred from the ΔK statistic of Evanno *et al*. (2005), denoted by $\widehat{K}_E$. It was also inferred from $\Pr[X|K]$ (the probability of obtaining the genotype data *X* given *K*) of Pritchard *et al*. (2000), denoted by $\widehat{K}_P$. For calculating $\widehat{K}_P$, the mean value of $\Pr[X|K]$ across the 20 replicate runs for each assumed $\widehat{K}$ value was obtained, and the $\widehat{K}$ value that had the highest mean $\Pr[X|K]$ was returned as $\widehat{K}_P$. In evaluating the qualities of $\widehat{K}_E$ and $\widehat{K}_P$, I was concerned only with the accuracy (below) and bias, not with the exact value when the estimate is different from *K*.

Accuracy assessments: For the inference of the number of populations (*K*) represented by a sample of individuals, the estimate $\widehat{K}_E$ or $\widehat{K}_P$ was compared with the actual (simulated) value of *K*, and the estimation accuracy was measured by the proportion of replicated datasets in which the estimate was equal to *K*. For a number of *m* replicates simulated under a given parameter combination, the accuracy of estimator $\widehat{K}_Y$ was calculated by $\Pr(\widehat{K}_Y = K) = \frac{1}{m}\sum_{j=1}^{m}(\widehat{K}_{Yj} = K)$, where the estimator *Y=E* for the ΔK method (Evanno *et al*. 2005) and *Y=P* for the $\Pr[X|K]$ method (Pritchard *et al*. 2000).

Measuring the accuracy of individual ancestry assignments is more difficult, because of the symmetry of the clustering model (Pritchard *et al*. 2000; Stephens 2000). A sample of individuals can be assigned to *K* populations in *K*! (labelling) ways with exactly the same likelihood, and the same biological meaning. For example, 4 individuals A, B, C and D can be assigned to *K*=3 populations, indexed by 1 to 3, in 6 equivalent partitions in which one population is represented by both A and B while the other two populations are represented by C and D respectively. These are {{A,B}, {C}, {D}}; {{A,B}, {D}, {C}}; {{C}, {A,B}, {D}}; {{C}, {D}, {A,B}}; {{D}, {A,B}, {C}}; {{D}, {C}, {A,B}}. In each partition, the *j*th (*j*=1, 2, 3) set of individuals is assigned to population *j*. Here population labels, 1 to 3, are arbitrary, and the 6 partitions have the same likelihood and the same biological meaning. At most, the MCMC algorithm of Structure returns one of the *K*! equivalent partitions (Pritchard *et al*. 2000). Different runs of the same data with different random number seeds and/or starting points may land on different ones of the *K*! equivalent partitions.

As the labels of the inferred populations are insignificant, I calculate the difference between the simulated (known) and estimated coancestry for pairs of individuals in a sample to measure the quality of individual ancestry inferences. For a total number of $n_0 = \sum_{j=1}^{K} n_j$ individuals sampled from *K* populations, the average assignment error, AAE, is measured by

$$AAE = \left(\frac{1}{n_0(n_0-1)/2} \sum_{i=1}^{n_0} \sum_{i'=i+1}^{n_0} \left(\sum_{j=1}^{\widehat{K}} \hat{q}_j^{(i)} \hat{q}_j^{(i')} - \sum_{j=1}^{K} q_j^{(i)} q_j^{(i')} \right)^2 \right)^{1/2}$$

where K and $\widehat{K}$ are the simulated (or known) and assumed numbers of populations, $q_j^{(i)}$ and $\hat{q}_j^{(i)}$ are the simulated (or known) and estimated ancestry of individual i (i=1, 2, …, n_0) coming from population j (j = 1~K and j = 1~$\widehat{K}$ for $q_j^{(i)}$ and $\hat{q}_j^{(i)}$ respectively), and $q_j^{(i')}$ and $\hat{q}_j^{(i')}$ are similarly defined for another individual i'. Obviously, *AAE* is invariable with label switching. Its minimum value is 0, when ancestry assignment is perfect (i.e. $\widehat{K}$=K, $\hat{q}_j^{(i)} = q_j^{(i)}$ for i=1~n_0 and j=1~K). Its maximum value is 1, when individuals from different populations are assigned to the same population and individuals from the same population are assigned to different populations. *AAE* was calculated for each of 20 runs of each of m=100 replicate datasets simulated with a given parameter combination, and the average across runs and replicates was reported.

An empirical dataset: To demonstrate the impact of unbalanced sampling and the effect of ancestry priors on Structure analysis in practice, I analysed a subset of the human data published in Rosenberg *et al*. (2005). The subset consists of 51 Palestinian individuals from Israel, 13 Colombian individuals from Colombia, and 24 Mandenka individuals from Senegal. Each individual was genotyped at 783 microsatellite loci. The three populations are well differentiated, and the sample of 88 individuals can be easily clustered into the 3 source populations (Palestinian, Colombian and Mandenka) with little admixture by Structure using genotype data only, no matter which (the default or the alternative) ancestry prior is used (Fig. S1, Supporting Information).

To investigate the impact of unbalanced sampling and ancestry prior, I generated subsamples by bootstrapping over individuals and over loci. Each subsample was obtained by keeping the original 51 Palestinian individuals, and by drawing at random (without replacement) 5 individuals from the 13 Colombian and 5 individuals from the 24 Mandenka. The genotypes of each individual in a subsample were obtained at a number of L (=10, 20, 40, 80, 160, 320, 640) loci, drawn at random (without replacement) from the original 783 microsatellites. For each L value, 200 subsamples were generated by the above procedure of bootstrapping over individuals and loci. Each subsample was then analysed by Structure using the admixture and correlated allele frequency models, a burn-in and running length of 10^4 iterations, and the default or the alternative ancestry priors. The assumed K value varied in the range [1, 5], and 20 replicate runs were conducted for each assumed K value. The most likely number of populations represented by the sampled 61 individuals was estimated by $\widehat{K}_E$ and $\widehat{K}_P$ as for the simulated data. The accuracy of $\widehat{K}_E$ and $\widehat{K}_P$ was

calculated by $P(\widehat{K}_E = K)$ and $P(\widehat{K}_P = K)$ respectively, where K=3. The assignment errors were calculated as *AAE*, assuming no individual has mixed ancestries as shown by the whole data (88 individuals, 783 microsatellites) analysis (Fig. S1, Supporting Information) and assuming a number of K=3 populations.

Results

Ancestry assignments: The default ancestry prior works well and yields accurate individual assignments (Fig. 1, 2) only when sampling is balanced (i.e. when n_1/n_i is close to 1, where $i \geq 2$). However, when sampling is unbalanced with roughly $n_1/n_i > 5$ in the case of K=3 (Fig. 1), individuals in the small samples from populations 2 and 3 are frequently assigned to a single cluster while individuals in the big sample from population 1 are frequently split into two or more clusters (see Fig. S2, Supporting Information). Similarly, individual ancestry assignments are also very poor for larger K values under unbalanced sampling (Fig. 2). An increasing extent of population differentiation measured by F_{ST} (Fig. S3, Supporting Information) and an increasing amount of marker information (Fig. 1) do not improve much of the assignment quality under this prior.

In contrast, the alternative prior yields highly accurate individual ancestry assignments (Fig. 1, 2; Fig. S3, Supporting Information), except when sampling is highly unbalanced and when either the actual number of populations is large (K>12, Fig. 2), the number of markers (L =10) is small (Fig. 1) or F_{ST} is low (0.05) (Fig. S3). In the case of a small number of populations (K=3, 6), almost perfect assignments were obtained by using the alternative prior, even when the larger sample is 38 times larger than the smaller samples (Fig. 1, 2). However, the superiority of the alternative prior decreases with an increasing number of populations, K (Fig. 2). When K is very large (≥24), the choice of prior has little effect on individual assignment accuracy, which is predominantly determined by the imbalance of sample sizes among populations (Fig. 2). An examination of assignment results shows that frequently the population represented by the large sample is split into two or more populations, while the populations represented by small samples are accurately reconstructed (Fig. S4, Supporting Information) under both priors. With many populations (K=48), the default prior performs slightly better than the alternative prior (Fig 2).

The poor performance of Structure when K is large and sampling is unbalanced (Fig 2) is caused mainly by its default ALPHA value. This default initial α value, ALPHA=1.0, seems to be too large that it impedes the mixing of the MCMC sampler to move to lower α values («1) which encourage the inference that each individual's ancestry comes mostly from a single population (i.e.

no admixture). With an increasing *K* and an increasing imbalance in sampling, the problem becomes increasingly severe, as is clear from Fig 3. Under the alternative ancestry prior, the individual assignments to populations improve quickly with a decreasing ALPHA value (Fig 3). This is true especially when the uncorrelated allele frequency model is used. In contrast, the default prior always leads to poor individual assignments to populations, regardless of the adopted allele frequency models and the values of ALPHA (Fig 3). Similar results were obtained for even larger *K* values (*K*=48). It can be concluded that the alternative ancestry prior, a non-default ALPHA value of about 1/*K*, and the non-default uncorrelated allele frequency model are required for Structure to give accurate inferences when *K* is large and sampling is unbalanced.

With balanced sampling (i.e. n_1/n_i close to 1), the alternative prior has the same assignment accuracy as the default prior. This is true for different numbers of markers (Fig. 1), different numbers of populations (Fig. 2) and different levels of population differentiation (Fig. S3, Supporting Information). There seems to be little accuracy cost of assuming the alternative prior in analysing data with balanced sampling.

Population number estimates: With balanced sampling, the number of populations represented by the sampled individuals, *K*, is estimated accurately by both estimators of $\widehat{K}_E$ and $\widehat{K}_P$, no matter which (default or alternative) prior of individual ancestry is used (Fig. 4; Fig. S3, Supporting Information). However, with an increasing difference in sample size between populations, both estimators deteriorate quickly, especially when the default prior was used. A close inspection of the results indicates that overestimates of *K* in some datasets and underestimates of *K* in other datasets were obtained by both $\widehat{K}_E$ and $\widehat{K}_P$ estimators. This is not surprising because, while the population represented by the large sample may be split, the populations represented by small samples may be merged in Structure analysis (Fig. S2, S4, Supporting Information). Most often $\widehat{K}_E$ under-estimates while $\widehat{K}_P$ over-estimates *K*.

Estimator $\widehat{K}_P$ under the alternative prior works noticeably better than the other estimator and prior combinations for the entire range of sample size differences (n_1/n_2 = 1~38) (Fig. 4; Fig. S3, Supporting Information). Overall, estimator $\widehat{K}_P$ is more accurate than $\widehat{K}_E$, no matter which (default or alternative) prior is used. While $\widehat{K}_E$ is reasonably accurate under balanced sampling, its performance deteriorates rapidly with increasing imbalance in sample sizes (Fig 4; Fig. S3, Supporting Information), and becomes highly inaccurate even when one sample is only 2 times larger than any of the other samples. The observations are consistent across different numbers of loci (Fig 4) and different differentiation levels (Fig. S3, Supporting Information).

The behaviour of $\widehat{K}_P$ under the default prior is complicated and perplexing. Its accuracy has two maxima when n_1/n_2 is about 1 and 10 respectively, and two minima when n_1/n_2 is about 5 and large, respectively (Fig. 4). This pattern is consistent across the cases of L=10, L=20 and L=40, and is thus unlikely due to insufficient replications. This erratic pattern is also observed cross different levels of population differentiation (Fig. S3, Supporting Information).

Empirical data analysis results: Better assignments with fewer assignment errors were obtained by using the alternative ancestry prior than the default ancestry prior (Fig. 5). The superiority of the alternative prior is substantial when the number of loci is small, and the assignment quality becomes independent of the priors when the number of loci is greater than 300.

Similar to the simulation results, $\widehat{K}_P$ is more accurate than $\widehat{K}_E$ when the number of loci is high (L > 40) (Fig. 5). The accuracy of $\widehat{K}_E$ is always smaller than 60%, no matter how many markers and which ancestry prior is used. In contrast, the accuracy of $\widehat{K}_P$ can reach a value as high as 85% when L > 80. It is puzzling that the accuracy of $\widehat{K}_P$ decreases gradually with an increasing L when L≥100 (Fig. 5). The alternative prior leads to a more accurate $\widehat{K}_P$ than the default prior when markers are numerous. These results confirm the conclusion from simulations that $\widehat{K}_P$ with the alternative prior provides the best estimate of K when sampling is unbalanced.

Discussion

My analyses of simulated and empirical datasets confirm the conclusions of previous studies (Kalinowski 2011; Neophytou 2014; Puechmaille 2016) that unbalanced sampling has a large impact on Structure analysis. It reduces the quality of individual assignments to populations, and the accuracy of the estimated number of populations using both the method ($\widehat{K}_E$) of Evanno *et al*. (2005) and the original method ($\widehat{K}_P$) of Pritchard *et al*. (2000). These effects become more severe with an increasing imbalance in size among samples from different populations, and cannot be removed by using many more markers (Fig. 1) or by increasing the levels of differentiation among populations (Fig. S3, Supporting Information). However, I showed that these adverse effects of unbalanced sampling on Structure analysis can be largely overcome by simply switching to the alternative ancestry prior, at least when the number of populations is not large. Under this alternative prior, each assumed population has a specific α value (which defines the prior proportional contribution of a population to the sample) which can be different from those of other populations. Under this alternative ancestry prior, Structure can yield highly accurate inferences of individual ancestries (Fig. 1, 2; Fig. S3) and reasonably good estimates of the number of

populations, *K*, (Fig. 4) represented by the sampled individuals, when sampling is unbalanced and *K* is not large.

In their simulations, Kalinowski (2011), Neophytou (2014), and Puechmaille (2016) considered 4, 3 and 10 populations, respectively. They showed that Structure yielded inaccurate inferences when sampling was unbalanced, as confirmed by the present study. This is mainly because the default ancestry prior was used in their Structure analyses. If the alternative prior, which is apparently more suitable for these unbalanced sampling situations, were used, the individual assignments to populations and the inference of *K* should be much improved, as demonstrated by the present study (Fig. 1~4).

Previous simulations have not considered many (say, $K>20$) populations. I showed that in the more difficult case of many populations, although Structure can still yield quality inferences of individual ancestry and *K* when sampling is balanced, it gives rather poor inferences when sampling is highly unbalanced, no matter which ancestry prior (default or alternative) is used, how differentiated the populations are, and how much marker information is available. My simulations demonstrate that further parameter and model adjustments in additional to individual ancestry prior are necessary for Structure to deal with this difficult situation of many populations under unbalanced sampling. Reducing the default ALPHA value (1.0) to approximately $1/K$ could improve the inferences substantially under the alternative prior (Fig 3). Adopting the best combination of the alternative prior, ALPHA ~ $1/K$, and the uncorrelated allele frequency model, Structure can yield highly accurate individual assignments even when *K* is extremely large (48) and sampling is highly unbalanced (Fig 3). These results imply that caution needs to be exercised about the default parameter settings in Structure, especially in difficult situations such as many populations and unbalanced sampling. In reality, the parameter *K* and the sample structure (whether balanced or not) for a dataset are unknown and are parts of the Structure inferences. It is therefore impossible to determine, *a priori*, the most suitable parameter settings for conducting a Structure analysis. I suggest, therefore, using multiple exploratory parameter combinations to make a comparative Structure analysis of the same dataset. These combinations should include the one with the alternative ancestry prior, an ALPHA value much smaller than the default (1.0) (say, ~ $1/K$ where *K* is the assumed number of populations), and the uncorrelated allele frequency model.

While the alternative prior can greatly improve individual assignments to populations (Fig. 1, 2, 3, 5; Fig. S3, Supporting Information), it does not help a lot in improving *K* inference (Fig. 4, 5; Fig. S3, Supporting Information). With balanced sampling, both $\widehat{K}_E$ and $\widehat{K}_P$ are accurate. However, the accuracy of both estimators deteriorates quickly with an increasing level of unbalanced

sampling (Fig. 4; Fig. S3, Supporting Information). When the sample size ratio n_1/n_2 is far from 1, both methods have a low frequency of recovering the actual number of populations. This is true regardless of individual ancestry priors, marker information contents and population differentiation levels. Relatively, $\widehat{K}_P$ outperforms $\widehat{K}_E$ substantially when sampling is unbalanced, and $\widehat{K}_P$ under the alternative prior yields the most accurate estimates of *K* (Fig. 4; Fig. S3, Supporting Information) at different levels of unbalanced sampling.

The simulation results highlight the difficulties of inferring *K* by Structure, as stressed by the authors (Pritchard *et al*. 2000) of the program. It is understandable that two or more optimal *K* values may exist to explain the genetic structure represented by a sample of individuals. Such a case occurs when samples are taken from hierarchically structured populations (Evanno *et al*. 2005), where different optimal *K* values may correspond to the numbers of populations defined at different hierarchical levels (e.g. continents, regions, sub-regions, …). It is also understandable that *K* may be arbitrary, depending on the sampling scheme. This is true when samples are taken from a large continuous population in Wright's (1946) neighbourhood migration model. However, my simulations considered a number of *K* well and equally differentiated populations in an island model, such that a single best *K* value should exist. Indeed, the frequency that *K*=3 is correctly recovered by both $\widehat{K}_P$ and $\widehat{K}_E$ estimators can reach 100% when sampling is balanced (Fig. 4), but reduces to about 20% when sampling is highly unbalanced. It is unclear why *K* is so poorly estimated while individual assignments to populations can be fairly accurate, when the alternative prior is applied to highly unbalanced sampling and when marker information and population differentiation are high.

My simulation and empirical data analyses show that $\widehat{K}_P$ is more accurate than $\widehat{K}_E$ when the sizes of samples from different populations are highly variable, and has a high accuracy similar to that of $\widehat{K}_E$ otherwise. In practice, however, the most widely applied estimator is $\widehat{K}_E$, being used thousands of times in published studies (Puechmaille *et al*. 2016). My results suggest that $\widehat{K}_P$ is preferable to $\widehat{K}_E$ in the simple situation of the island migration model where all populations are differentiated more or less equally without a hierarchical structure. However, more simulations considering more complicated situations, such as hierarchically structured populations (Evanno *et al*. 2005), populations differentiated to different degrees from the ancestral (or pooled) population in the otherwise island model, are needed to compare $\widehat{K}_P$ and $\widehat{K}_E$ before a general conclusion can be reached. In the meantime, I suggest use both estimators in inferring *K* in practice.

A survey of published studies shows that none explicitly stated that the alternative ancestry prior and a non-default ALPHA value were used in Structure analysis. Frequently the default

parameter settings of Structure, including the default ancestry prior and ALPHA=1.0, were adopted without justification. This is not surprising, as the genetic model of Structure is complicated with many parameters requiring initial values (like ALPHA=1.0), prior distributions (like prior individual ancestry) and model selection (like correlated and uncorrelated allele frequency models), and the choice of some parameter settings is not always obvious. With almost every parameter having a default value/prior which is easily accepted by users without a second thought, it is all too easy to run Structure analysis. It is however also easier to misuse Structure because the default parameter settings may not be suitable for a particular dataset. Given the ubiquity of unbalanced sampling, and the large advantage and low cost of the alternative (population specific) ancestry prior as demonstrated in this study, I suggest the authors of Structure to set the population specific ancestry prior and ALPHA=1/*K* as the defaults, where *K* is the assumed number of populations in a particular Structure analysis. I also suggest users of the program to choose the population specific ancestry prior and ALPHA=1/*K* in analysing their data, and to infer *K* using both $\widehat{K}_P$ and $\widehat{K}_E$. I further suggest the authors and users of Structure to pay more attention to the uncorrelated allele frequency model, although it can be argued that population allele frequencies are usually correlated as measured and modelled by F_{ST}. At a minimum, this model should be explored together with the alternative ancestry prior and an ALPHA value much smaller than 1.0 when the default parameter setting in Structure does not work well for a dataset. The most likely number of populations estimated by Structure should be taken with caution, especially when sampling is suspected to be unbalanced or/and the number of populations is large.

Literature

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611-2620.

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.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574-578.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322-1332.

Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity*, **106**, 625-632.

Neophytou C (2014) Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. *Tree Genetics & Genomes*, **10**, 273-285.

Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefansson K, *et al*. (2002) Assessing population differentiation and isolation from single nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B*, **64**, 695–715

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu M (2013) An overview of STRUCTURE: applications, parameter settings and supporting software. *Frontiers in Genetics*, **4**, 98.

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

Puechmaille SJ (2016) The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources*, **16**, 608-627.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, **73**, 1402-1422.

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**, 660-671.

Smouse PE, Waples RS, Tworek JA (1990) A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 620-634.

Stephens M (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795-809.

Wang, J., 2006. Informativeness of genetic markers for pairwise relationship and relatedness inference. *Theoretical Population Biology*, **70**, 300-321.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39-59.

Data Accessibility

Source code and (Windows) executable for simulating genotype data, preparing input files for Structure, and running Structure: DRYAD entry DOI: http://dx.doi.org/10.5061/dryad.f8n5j

Source code and (Windows) executable for generating bootstrapping subsamples of the human dataset, preparing input files for Structure, and running Structure: DRYAD entry DOI: http://dx.doi.org/10.5061/dryad.f8n5j

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The inferred membership of 88 human individuals in K=3 populations by Structure. Each individual was genotyped at 783 microsatellites. The results were obtained from Structure using a burn-in length and a run length of 10^4 iterations, the admixture and correlated allele frequency models, the default (upper panel) or the alternative (lower panel) ancestry prior, and all other parameters at default values. Each individual is represented by a thin line partitioned into K coloured segments that represent the individual's estimated membership fractions in K populations.

Fig. S2 The inferred membership of individuals in an example simulated dataset by Structure. The dataset has 200 simulated individuals drawn from K=3 populations in the island model at drift-migration-mutation equilibrium with F_{ST}=0.10. Individuals 1~180, 181~190, and 191~200 (ordered from left to right on the x axis) were sampled from populations 1, 2, and 3, respectively, and each individual was genotyped at 40 loci with each having 10 alleles. The analysis results were obtained from Structure using a burn-in length and a run length of 10^4 iterations, the admixture and correlated allele frequency models, the default (upper panel) or the alternative (lower panel) ancestry prior, and all other parameters at default values. Each individual (on the x axis) is represented by a thin line partitioned into K=3 coloured segments that represent the individual's estimated membership fractions in K populations.

Fig. S3 Quality of individual assignments to populations and inferences of K by Structure. Three equally differentiated populations (K =3) with F_{ST}=0.05, 0.10 or 0.20 in the island model were simulated. Each individual has genotypes at L=20 loci, each having 10 alleles. The sample size was

n_i for population i (i=1, 2, 3), with $n_2 \equiv n_3$ and $n_1 + n_2 + n_3 \equiv 200$. At each F_{ST} value, 100 replicate datasets were analysed by Structure using the default (Dft) and alternative (Alt) priors. The inference quality was measured by average assignment errors (*AAE*) in panel A, and by $P(\widehat{K}_E = K)$ and $P(\widehat{K}_P = K)$ in panels B, C, and D, as a function of the extent of unbalanced sampling measured by the sample size ratio, n_1/n_2 (x axis). A burn-in and run length of 10^4, the admixture and correlated allele frequency models, and the default values of other parameters were used in Structure analyses.

Fig. S4 The inferred membership of an example simulated dataset of 1600 individuals from K=24 populations by Structure. Individuals 1~680, 681~720, 721~760, …, 1561~1600 (ordered from left to right on x axis) were sampled from populations 1, 2, 3, …, 24 respectively, the sample size being 680 for the first population, and 40 for each of the other 23 populations. Each individual was genotyped at 50 loci with each having 10 alleles. The populations were at drift-migration-mutation equilibrium with F_{ST}=0.20. The analysis results were obtained from Structure using a burn-in length 5×10^5 and a run length 10^4 iterations, the admixture and correlated allele frequency models, the default or the alternative ancestry prior, and all other parameters at default values. Each individual (on the x axis) is represented by a thin line partitioned into K=24 coloured segments that represent the individual's estimated membership fractions in K populations.