A parsimony estimator of the number of populations from a STRUCTURE-like analysis

Jinliang Wang

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

Left running head: J Wang

Right running head: Estimating the number of populations

Key words: STRUCTURE, markers, genetic differentiation, number of populations

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

Population genetics model based Bayesian methods have been proposed and widely applied to making unsupervised inference of population structure from a sample of multilocus genotypes. Usually they provide good estimates of the ancestry (or population membership) of sampled individuals by clustering them probabilistically or proportionally into (anonymous) populations. However, they have difficulties in accurately estimating the number of populations (K) represented by the sampled individuals. This study proposed a new *ad hoc* estimator of K, calculable from the output of a population clustering program such as STRUCTURE or ADMIXTURE. The new criterion, called parsimony index (PI), aims to identify the number of populations (K) which yields consistently the minimal admixture estimates of sampled individuals. Extensive simulated and empirical data were used to compare the accuracy of PI and two popular K estimators based on $\Pr[X|K]$ (i.e. the probability of genotype data X given K) and ΔK (i.e. the rate of change of the probability of data as a function of K) calculated from STRUCTURE outputs, and the accuracy of PI and the cross-validation method calculated from ADMIXTURE outputs. It was shown that PI was more accurate than the other methods consistently in various population structure (e.g. hierarchical island model, different extents of differentiation) and sampling (e.g. unbalanced sample sizes, different marker information contents) scenarios. The ΔK method was more accurate than the $\Pr[X|K]$ method only for hierarchically structured or highly inbred populations, and the opposite was true in the other scenarios. The PI method was implemented in a computer program, KFinder, which can be run on all major computer platforms.

Introduction

Traditional population structure analysis uses Wright's (1951) F statistics to describe and understand the patterns of genetic variation in populations. These statistics, F_{IS}, F_{ST} and F_{IT}, can be estimated from genetic marker data collected from several populations (e.g. Weir and Cockerham 1984), and offer a convenient and elegant means of summarising population structures. However, the analysis of F statistics relies on information of known predefined (e.g. by geographic locations) source populations of sampled individuals. In reality, however, the information might be unavailable, incomplete, or unreliable for sorting individuals into populations. A typical example is mixed stock analysis (Smouse *et al.* 1990), where

individuals coming from different source populations to mix in the same feeding/breeding ground are sampled to determine the genetic structure. Another typical example is a batch of seized illegally traded animals or animal parts (e.g. tusks). In both examples, the source populations of the sampled individuals are unknown and are the primary interest of analysis (Hsieh *et al*. 2003; Velo-Anton *et al*. 2007).

Pritchard *et al*. (2000) proposed a Bayesian method, based on a population genetics model, to identify populations represented by a sample of individuals and to assign the individuals (or their genomes) probabilistically to the identified source populations using individual genotypic data. This is possible because population genetics theory tells us that individuals from the same source population share the same gene pool, and thus have similar multilocus genotypes that are roughly in Hardy-Weinberg equilibrium and linkage equilibrium. The method, implemented in the computer program STRUCTURE, has largely solved the problem challenging the traditional population structure analysis, and has revolutionized our ability to conduct unsupervised population structure analysis using marker genotypes only. Following Pritchard *et al*. (2000), many similar methods (e.g. Dawson and Belkhir 2001; Dupanloup *et al*. 2002; Corander *et al*. 2003; Guillot *et al*. 2005; Tang *et al.* 2005; François *et al.* 2006; Gao *et al.* 2007; Huelsenbeck and Andolfatto 2007; Alexander *et al.* 2009; Jombart *et al.* 2010; Raj *et al.* 2014) have been developed to infer population structure with higher computational efficiency and with extended models (e.g. inbreeding models to accommodate inbred individuals and spatial models to use geographic as well as genetic data). The most popular method remains that of Pritchard *et al.* (2000), because of its accurate, robust and versatile models thanks to Pritchard and coworker's original and continued work (e.g. Falush *et al.* 2003; 2007; Hubisz *et al.* 2009). The dominance of STRUCTURE over other programs in marker-based population structure/admixture analyses, even in this genomics era, is readily confirmed by a survey of studies published in the most recent issues of peer-reviewed journals such as Molecular Ecology.

STRUCTURE and related methods work well in assigning individuals to their source populations for a given number of populations, *K*. When population differentiation is substantial or/and marker information is sufficient, they give accurate individual ancestry (or population membership) inferences. However, they have difficulties in identifying source populations and inferring the optimal number of populations, *K*, represented by the sampled individuals. Pritchard *et al.* (2000) proposed an *ad hoc* procedure to estimate the marginal likelihood Pr[*X*|*K*], the probability of obtaining the genotype data *X* given *K*. The *K* value that

maximizes Pr[$X|K$] is the best estimate of the number of populations. They demonstrated, using a couple of simple empirical datasets, that the method works well. Evanno *et al.* (2005) found by simulations that the Pr[$X|K$] method gives poor estimates of K for hierarchically structured populations, and proposed another *ad hoc* statistic, ΔK (i.e. the rate of change of the probability of data as a function of K), to estimate the number of populations at the uppermost hierarchical level of structure. Alexander and Lange (2011) employed a cross-validation method, implemented in ADMIXTURE software, to identify the best K value as judged by the prediction of systematically withheld data points. Gao *et al*. (2007) and Durand *et al*. (2009) used the deviance information criterion (DIC) for inferring K in InStruct and TESS programs, respectively. Dawson and Belkhir (2001), Pella and Masuda (2006), and Huelsenbeck and Andolfatto (2007) took both K and individual assignments to populations as random variables and used joint priors, such as a Dirichlet process prior, to estimate both in programs PARTITION and STRUCTURAMA. Corander *et al*. (2003, 2004) implemented a split-and-merge algorithm in their program BAPS to estimate K. Patterson *et al*. (2006) proposed an eigenanalysis method, implemented in SmartPCA software, to estimate K as 1 plus the number of significant eigenvalues explaining the variation of genotype data. Jombart *et al.* (2010) and Beugin *et al.* (2018) used Akaike information criterion (AIC, Akaike 1998), Bayesian Information Criterion (BIC, Schwarz 1978), Kullback Information Criterion (KIC, Cavanaugh 1999) and their variants to assess the best supported model, and therefore the most likely number of populations. These and other methods were demonstrated to yield good estimates of K in some simple scenarios (e.g. Gao *et al*. 2011), but can be highly inaccurate in difficult situations such as many source populations (say, K>10), unbalanced sample sizes (Wang 2017), hierarchical population structures (Evanno *et al.* 2005), weak differentiation or low marker information (Gao *et al*. 2011), and high admixture.

Except for the cross-validation method (Alexander and Lange 2011) and the significant eigenvalue method (Patterson *et al*. 2006), all K estimators described above are based (in one form or another) on the estimated marginal likelihood of the model or the probability of data. This is, in theory, a natural choice for measuring model fit. In practice, however, several problems arise. First, this quantity is difficult to calculate accurately, and as a result, some *ad hoc* approximation is adopted (Pritchard *et al.* 2000). It is unclear how well the approximation works, especially when marker information is insufficient or the inference is difficult (e.g. with many populations, unbalanced sampling, and low differentiation). Second, the likelihood maximization procedures (e.g. expectation maximization algorithm,

EM) or the Bayesian Markov Chain Monte Carlo (MCMC) procedures may not converge for this high dimensional optimization problem. The number of parameters to be estimated are roughly $N(K-1)+LK(A-1)$, where N, K, A, and L are the number of sampled individuals, number of populations, average number of alleles at a locus, and the number of loci. The estimated likelihood or probability of data could vary substantially among replicate runs of the same data, especially in difficult situations (above). Third, the criteria of these *ad hoc* methods in selecting the best K value may not be appropriate. For example, AIC, BIC, and KIC are all based on the same principle, assessing model quality by considering its likelihood against its complexity. Apparently, the penalty for model complexity is different among these criteria, and it is unclear which (if any) is the most suitable for this clustering problem. It is possible that none applies in general, and some modified forms of these criteria (e.g. Chen and Chen 2008; Gao and Song 2010) might be more appropriate for this high dimensional clustering problem.

In this paper, I propose another *ad hoc* criterion to estimate K, and use extensive simulations and empirical data to show that it is in general more accurate than previous methods. The criterion is based mainly on the quality of individual ancestry estimates from STRUCTURE-like programs, and the best K is the one that consistently yields the minimal mean admixture of sampled individuals. This principle, called minimal admixture or parsimony for simplicity hereafter, is derived from the observation that suboptimal K (i.e. values higher or lower than the true K) usually leads to inconsistent and inflated admixture (mixed ancestry) estimates from Bayesian (e.g. Pritchard *et al.* 2000) or likelihood (e.g. Tang *et al.* 2005; Alexander *et al.* 2009) population clustering analyses. The method is implemented in a computer program, KFinder (https://www.zsl.org/science/software/KFinder), to yield the best K given the outputs for a range of K values from a STRUCTURE-like program. Extensive simulations, considering many population scenarios (e.g. differentiation levels, subdivision models) and sampling scenarios (e.g. sample sizes per subpopulation, numbers of markers), were conducted to compare the performances of the parsimony method, the $\Pr[X|K]$ method (Pritchard *et al*. 2000) and the ΔK method (Evanno *et al.* 2005) in estimating K from STRUCTURE outputs, and of the parsimony method and the cross-validation method (Alexander and Lange 2011) in estimating K from ADMIXTURE outputs. I showed that the parsimony method improves K estimates consistently and sometimes dramatically over other methods calculated from both programs STRUCTURE and ADMIXTURE.

Methods

The parsimony method

I assume a dataset was analysed by STRUCTURE (or related methods such as ADMIXTURE) under the same set of model parameters (e.g. admixture and correlated allele frequency models) except for different K values, from a low bound K_L to a high bound K_H. For each K value ($K = K_L, K_{L+1}, \ldots, K_{H-1}, K_H$), a number of n_r =20 replicate runs were conducted, following Evanno *et al.* (2005). Therefore, the total number of runs (and output files) for a single dataset is $N_R = (K_H - K_L + 1)\, n_r$. Parameter options for STRUCTURE and ADMIXTURE runs are detailed below.

A parsimony index was calculated for each K in the range K_L to K_H, using the information in the n_r replicate-run output files. First, the mean estimate of the log probability of data, Pr[X|K], was calculated from the n_r output files, and the replicate runs with Pr[X|K] values smaller than the mean were discarded from further analyses. This is because the clustering algorithms may not converge, and different runs may end up with highly different Pr[X|K] values and individual ancestry estimates. Replicate runs with low Pr[X|K] values are expected to give poor estimates of individual ancestries and are thus abandoned. Note in the best scenario where all n_r replicate runs converge, no runs are abandoned, such that the number of retained runs, $n_r' = n_r$.

Second, an assignment quality score is calculated for each of the n_r' retained replicate runs. The (main) source population of an individual is determined as the one that has the largest ancestry coefficient estimate for the individual. The average co-assignment score within populations is calculated as

$$ACS_w = \frac{1}{n_w} \sum_{i=1}^{N} \sum_{j=1, ij \in \Phi}^{i} \sum_{l=1}^{k} q_l^{(i)}\, q_l^{(j)}, \qquad (1)$$

where $q_l^{(i)}$ and $q_l^{(j)}$ are the estimated proportions of individual i's and individual j's genomes that originate from population l (=1, 2, …, k), respectively, $ij \in \Phi$ signifies that individuals i and j are inferred to come from the same (main) source population, n_w is the number of pairs of individuals (including an individual with itself) that share the same inferred source population, and N is the number of sampled individuals. The quantity calculated by (1) is similar to Dawson and Belkhir's (2001) probability of co-assignment. Both measure the probability that a set of individuals belong to the same population, independent of the arbitrary labelling of source populations.

The average co-assignment score between populations, ACS_b, is calculated similarly by (1), except n_w is replaced by n_b, the number of pairs of individuals that have different source populations, and $ij \in \Phi$ is replaced by $ij \in \emptyset$ which signifies that individuals *i* and *j* are inferred to belong to different source populations.

` The strength of population structure is characterized by a high value of ACS_w close to 1, and a small value of ACS_b close to 0. When there is no hybridization (admixture) and individual ancestry inference is perfect, ACS_w reaches its maximal value of 1 and ACS_b reaches its minimal value of 0. An overall measure of the strength of population structure is

$$SPS = ACS_w - ACS_b. \quad (2)$$

Third, the harmonic mean of the sizes of well-defined clusters is calculated. For each individual *i* (=1, 2, …, *N*), its main source population is determined to be *l* (=1, 2, …, *k*) if $q_l^{(i)}$ is the largest and $q_l^{(i)} \geq Q_{min}$, where Q_{min} is a chosen threshold value (say, 0.8). Its main source population is undetermined if $q_l^{(i)} < Q_{min}$. The proportion of the *N* sampled individuals whose main source population can be determined is then calculated. If this proportion is not smaller than Q_{min}, then main cluster structure of the sample is obtained. Otherwise, the value of Q_{min} is halved and used to repeat the above process until the main cluster structure of the sample is attained. The size of the *l*th (for *l*=1, 2, …, *k*) cluster, S_l, is calculated as the number of individuals whose main source populations are determined to be *l*. The harmonic mean of the cluster sizes that are larger than 5, *HMCS*, is then calculated,

$$HMCS = n_{5+} / \sum_{l=1, S_l>5}^{k} S_l^{-1}, \quad (3)$$

where n_{5+} is the number of clusters with each containing 5 or more individuals. The threshold cluster size, 5, is more or less arbitrary. However, it is chosen to reduce the population splitting errors and the population merging errors.

Fourth, the overall strength of the inferred population structure for a given run is calculated as

$$SPS' = \frac{SPS}{HMCS}. \quad (4)$$

Among the n'_r runs retained after Pr[*X*|*K*] screening in step 1, the largest value of SPS', SPS^*, is obtained.

Fifth, the total number of clusters whose sizes are not larger than 5 (as determined in step 3) across the n'_r retained runs, n_{5-}, is calculated.

Sixth, the consistency of ancestry assignments cross runs is calculated as

$$CAA = \frac{1}{N(N-1)/2}\sum_{i=1}^{N}\sum_{j=1}^{i-1}(\delta_{0,m}^{(ij)} + \delta_{n'_r,m}^{(ij)}), \quad (5)$$

where m is the number of runs among a total number of n'_r retained ones in which individuals i and j are assigned to the same cluster, $\delta_{0,m}^{(ij)} = 1$ and 0 if m=0 and $m > 0$ respectively, $\delta_{n'_r,m}^{(ij)} = 1$ and 0 if m=n'_r and $m < n'_r$ respectively. *CAA* measures the consistency in main cluster assignments among replicate runs. Its maximal and minimal values are 1 and 0, respectively.

Seventh, the overall assignment quality for an assumed number of k populations is measured by the parsimony index

$$PI = SPS^* + CAA - \frac{2n_{5-}}{kn_r} - \frac{1}{2k}. \quad (6)$$

The last term in (6) is a penalty against small k because both *SPS* and CAA tend to increase with a decreasing k value. In the extreme case of k=1, $SPS^* \equiv 1$ and $CAA \equiv 1$ because all N individuals must be inferred to come from the same source population (i.e. $ACS_w \equiv 1$ and $ACS_b \equiv 0$ in (2) and $\delta_{0,m}^{(ij)} \equiv 0$ and $\delta_{n'_r,m}^{(ij)} \equiv 1$ in (5)).

For each k in the range K_L to K_H, a corresponding value of PI is calculated. The k value that yields the largest PI value is inferred to be the most likely number of populations represented by the sample, K.

Simulations

Simulated data were generated under different population models and sampling intensities, and used to evaluate the accuracy of the above described parsimony index (PI) and two popular methods, Pr[X|K] (Pritchard *et al.* 2000) and ΔK (Evanno *et al.* 2005), in estimating K from STRUCTURE. Because increasingly large SNP datasets are produced and analysed by ADMIXTURE and other programs faster than STRUCTURE, I also simulated data with many SNPs and estimated K using PI and the cross-validation method (Alexander and Lange 2011) calculated from ADMIXTURE outputs.

I assumed Wright's (1931) island (IS) model or a two-level hierarchical island (HI) model (Evanno *et al.* 2005) for population structure in simulating genotype data. A number of N_k individuals were drawn at random from population k (k = 1, 2, . . ., K), and each sampled individual was genotyped at a number of L loci, each having A codominant alleles.

The ancestral allele frequencies at a marker locus l (l=1, 2, …, L), $\boldsymbol{p_{0l}} = \{p_{0l1}, p_{0l2}, \ldots, p_{0lA}\}$, were drawn from a uniform Dirichlet distribution, $\mathcal{D}(1,1,\ldots,1)$. The corresponding allele frequencies of population i under IS model, $\boldsymbol{p_{il}} = \{p_{il1}, p_{il2}, \ldots, p_{ilA}\}$, were drawn from $\boldsymbol{p_{0l}}$ following the Dirichlet distribution $\mathcal{D}(fp_{0l1}, fp_{0l2}, \ldots, fp_{0lA})$, where $f = 1/F_{ST} - 1$ (Nicholson *et al.* 2002; Falush *et al.* 2003) and F_{ST} was the genetic differentiation among populations. For HI model, the allele frequencies of an archipelago a, $\boldsymbol{p_{al}} = \{p_{al1}, p_{al2}, \ldots, p_{alA}\}$, were sampled from $\boldsymbol{p_{0l}}$ using $\mathcal{D}(fp_{0l1}, fp_{0l2}, \ldots, fp_{0lA})$, where $f = 1/F_{ST1} - 1$ and F_{ST1} was the genetic differentiation among archipelagos. The allele frequencies of an island i within an archipelago a were sampled from $\boldsymbol{p_{al}}$ using $\mathcal{D}(fp_{al1}, fp_{al2}, \ldots, fp_{alA})$, where $f = 1/F_{ST2} - 1$ and F_{ST2} was the genetic differentiation between islands within the archipelago.

Given the allele frequencies of a population (or island) i, the genotype of an individual sampled at random from the population at L loci were generated, assuming Hardy–Weinberg equilibrium and linkage equilibrium. The data, a number of N_k (k=1, 2, …, K) multilocus genotypes sampled from population k, were then pooled across populations and were subjected to STRUCTURE or/and ADMIXTURE analysis. Data simulated with a few loci and many alleles per locus were analysed by STRUCTURE only, while data simulated with many SNPs (A=2) were analysed by both STRUCTURE and ADMIXTURE, or by the latter only.

Simulation 1, differentiation F_{ST}: The accuracy of a population structure analysis relies on the strength, measured by F_{ST}, of the true structure. Populations of low F_{ST} values (close to 0) are difficult to identify and thus the number of populations represented by a sample of individuals is difficult to estimate from a STRUCTURE-like analysis. This simulation investigated the impact of F_{ST} in an IS model of K=6 populations on different K estimators. The 6 populations were assumed to differentiate from the ancestor population to the same extent of F_{ST} =0.02, 0.04, 0.08 or 0.16, or to different extents of F_{ST} =0.02 for populations 1 and 2, F_{ST} =0.04 for populations 3 and 4, and F_{ST} =0.08 for populations 5 and 6. Thirty individuals from each population were genotyped at 20 (equal F_{ST}) or 50 (unequal F_{ST}) loci, each having A=10 alleles.

Simulation 2, number of loci L: More markers provide more information and thus should yield more accurate inferences of population structure and *K*. This simulation considered an IS model of 6 populations differentiated to the same level of F_{ST} =0.1. Thirty individuals were sampled from each population and genotyped at a varying number of loci (each having *A*=10 alleles), *L*=4, 6, 8, 10, 12, 14, 16, 18, 20.

Simulation 3, number of populations K: It becomes increasingly challenging to estimate *K* accurately with an increasing number of populations represented by a sample of individuals. This simulation generated data from an IS model of *K* (=1, 2, …, 20) populations differentiated to the same level of F_{ST} =0.1, and compared the accuracy of different *K* estimators. Thirty individuals were sampled from each population and genotyped at *L*=10 and *L*=20 loci, each having *A*=10 alleles.

Simulation 4, unbalanced sampling: Population structure is difficult to infer from a sample containing many individuals from one population but few individuals from another. Heavily represented populations tend to split while lightly represented populations tend to merge in reconstructing the population structure from such an unbalanced sample of individuals. This simulation considered *K*=3 populations in an IS model with F_{ST} =0.1. The sample size was fixed at 300 individuals, with a number of *X* individuals sampled from population 1 and the remaining 300 – *X* individuals sampled equally from populations 2 and 3. *X* took values of 100, 120, 140, …, 280, resulting in a perfectly balanced sample when *X*=100, and a highly unbalanced sample when *X*=280. Each sampled individual was genotyped at *L*=20 loci, each having *A*=10 alleles.

Simulation 5, hierarchical structure: The HI model has two true *K* values, the number of archipelagos (K_a) and the number of islands (K_i). While the ΔK method (Evanno *et al.* 2005) was shown to estimate K_a, it is unclear what the Pr[*X*|*K*] method (Pritchard *et al.* 2000) estimates. Is it K_a, K_i, or neither? The PI method estimates K_i, because the islands have a much smaller harmonic mean cluster size than archipelagos. This simulation considered a HI model of K_a archipelagos, each containing K_a islands (such that $K_i = K_a^2$), where K_a =2, 3 and 4. Both F_{ST1} and F_{ST2} were assumed 0.1, and 30 individuals were sampled from each island (total sample size $N = 30K_a^2$). Each sampled individual was genotyped at *L*=20 loci, each having *A*=10 alleles.

Simulation 6, hybridization: PI index was partially based on minimizing the estimated admixture, and its accuracy might be compromised for a sample containing many hybrid

individuals. This simulation considered an IS model of K=3 populations with F_{ST}=0.1 and different degrees of hybridization. A sample contained 50 individuals from each population, among which a proportion H were either F1 or F2 hybrids (at equal probabilities) between the resident population and any of the other populations (with an equal probability). Each sampled individual was genotyped at L =20 loci, each having A=10 alleles.

Simulation 7, inbreeding: Inbreeding causes correlation between the homologous genes at a locus within an individual, and thus a loss of information in inferring population structures. This simulation considered different degrees of inbreeding (due to selfing) in an IS model of K=5 populations with F_{ST}=0.1. A sample contained 30 individuals from each population, each individual being produced by self-reproduction at a rate s (s=0, 0.05, 0.1, 0.2, 0.4, 0.8) or by outbreeding at a rate 1-s. Each sampled individual was genotyped at L =10 or 20 loci, each having A=10 alleles.

Simulation 8, many SNPs and low F_{ST}: With genomic data of many SNPs, it is now possible to infer population structure even when it is rather weak (i.e. F_{ST} small). This simulation considered an IS model of K=5 populations with F_{ST}=0.01. A sample of 20 individuals were drawn from each population, and each sampled individual was genotyped at L =100, 200, 400, 800, 1600, 3200, 6400, 12800, and 204800 loci, each having A=2 alleles. The data were analysed by both STRUCTURE (except for L=204800) and ADMIXTURE, and K was inferred by PI, Pr[X|K], ΔK and cross-validation methods.

Simulation 9, many SNPs and variable F_{ST}: Simulation 8 showed that the cross-validation method always inferred K=1, while the truth is K=5, even when L=204800 loci were used in ADMIXTURE analysis which yielded almost perfect population assignments under K=5. Simulation 9 was conducted to investigate whether the extremely poor performance of ADMIXTURE's cross-validation method was due to the low F_{ST} (0.01) or not. For this purpose, K=5 populations in the island model with variable F_{ST} values (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10) were simulated. A sample of 20 individuals were drawn from each population, and each sampled individual was genotyped at L =1000 loci, each having A=2 alleles. The data were analysed by ADMIXTURE, and K was inferred by the PI and cross-validation methods.

Structure analysis

The simulated data were analysed by STRUCTURE program (version 2.3.4, Pritchard *et al.* 2000), and the analysis results were further analysed by the Pr[X|K] method (Pritchard *et al.* 2000), the ΔK method (Evanno *et al.* 2005) and the new PI method for the most likely number of populations, K. The model and parameter settings adopted in the analyses were admixture model and correlated allele frequency model, INFERALPHA=1, ALPHA=1.0, POPALPHAS=1, UNIFPRIORALPHA=1, ALPHAMAX=10.0, ALPHAPROPSD=0.025. The alternative prior for individual ancestry was adopted, because it gave much more accurate STRUCTURE analysis results than the default prior when sampling was highly unbalanced (Wang 2017). The burn-in length was 10^4, 10^5 and 10^6 iterations in analysing data simulated with $K < 6$, $6 \leq K < 10$, and $K \geq 10$ populations, respectively. More populations lead to more parameters for STRUCTURE to estimate and therefore pose more challenges for the MCMC algorithm to converge. These burn-in lengths were obtained by experimenting with many pilot analyses of data with different simulated K values. The run length was 10^4 iterations. For all other parameters not mentioned above, their default values were used.

For a dataset simulated with a given K, STRUCTURE analyses were conducted for each assumed number of populations in the range [K_L, K_H], where K_L =Max[K-3, 1] and K_H =K+3. This narrow range was adopted to reduce the computational burden, and because the primary interest was whether the true simulated K value was recovered or not. For each assumed K value k, a number of n_r =20 replicate runs were conducted.

The simulated data with many SNPs were also analysed by ADMIXTURE program (version 1.3.0, Alexander and Lange 2011). The SNP data were first reformatted by PLINK (version 1.9.0, Purcell *et al*. 2007), and then analysed by ADMIXTURE using the program's default settings. For each dataset and each assumed K in the range [K_L, K_H] where K_L =Max[K-3, 1] and K_H =K+3, a number of 10 independent replicate runs of ADMIXTURE were conducted. The most likely K was inferred from ADMIXTURE outputs using its default cross-validation method and the PI method.

All simulations and data analyses described above were conducted on a large Linux cluster, using many cores in parallel by MPI.

Accuracy assessment

For each simulation (i.e. set of parameters), a number of 50 replicate datasets were generated. Each replicate dataset was analysed by STRUCTURE or/and ADMIXTURE assuming k in

the range [K_L, K_H], and was replicated with n_r =20 for each assumed K value. The total number of STRUCTURE-like analyses for a single dataset was thus $(K_H - K_L + 1) \times 20$, which was 80, 100, 120, 140 when the simulated K value was 1, 2, 3, and ≥4, respectively. The results in STRUCTURE's output files were analysed by the ΔK method as described by Evanno *et al.* (2005), the Pr[X|K] method (Pritchard *et al.* 2000) as described by Wang (2017), and the PI method as described above to obtain the estimates of K, denoted by $\widehat{K}_{Ev}$, $\widehat{K}_{Pr}$, and $\widehat{K}_{PI}$ respectively. The ADMIXTURE program outputs were used to yield the K estimates from both PI method and the cross-validation method (Alexander and Lange 2011), denoted by $\widehat{K}_{AL}$.

The accuracy of an estimator was measured by the proportion of replicate datasets in which the estimator was equal to the simulated true K. These accuracy measurements were denoted as $\Pr(\widehat{K}_{Ev} = K)$, $\Pr(\widehat{K}_{Pr} = K)$, $\Pr(\widehat{K}_{PI} = K)$ and $\Pr(\widehat{K}_{AL} = K)$ for estimators $\widehat{K}_{Ev}$, $\widehat{K}_{Pr}$, $\widehat{K}_{PI}$ and $\widehat{K}_{AL}$, respectively.

A human dataset

The performance of the three K estimators, $\widehat{K}_{Ev}$, $\widehat{K}_{Pr}$, and $\widehat{K}_{PI}$, was also compared by analysing a human dataset, published in Wang *et al*. (2007). The dataset contains 1484 individuals sampled from 78 world-wide populations, each individual being genotyped at 678 microsatellite loci. It proves to be difficult to reconstruct the population structure unambiguously from the 1484 sampled individuals, even using all of the 678 highly polymorphic microsatellites (Wang *et al*. 2007). For demonstration purpose and for reducing computational burden of a bootstrapping analysis, I choose to analyse a sub-dataset composing of 24 Basque individuals sampled from France, 19 Melanesian individuals from Bougainville, 21 Surui individuals from Brazil, 22 Mandenka individuals from Senegal, and 29 Japanese individuals from Japan. These populations are well differentiated and have balanced sample sizes, and as a result can be distinguished using about 20 markers. A number of 100 replicate datasets were generated by bootstrapping over loci, for L = 10, 20, 40, 80 and 160. Each dataset was analysed by 20 replicate STRUCTURE runs for each assumed K value from 1 to 10 (including the true value of K=5). The three K-estimators were then applied to the STRUCTURE outputs. Accuracy of the estimators was evaluated by calculating the proportions of the replicate datasets in which K<5, K=5, and K>5 were obtained.

Results

Simulation 1, differentiation F_{ST}

All three estimators become more accurate with an increasing F_{ST} (Figure 1). The accuracy of estimators $\widehat{K}_{Ev}$ and $\widehat{K}_{Pr}$ is similar, and is consistently lower than that of $\widehat{K}_{PI}$.

Estimator $\widehat{K}_{PI}$ has an accuracy increasing rapidly with an increasing F_{ST} value (Figure 1), and it recovers the simulated K value completely (i.e. accuracy = 100%) when $F_{ST} \geq 0.08$. The low accuracy at a small F_{ST} value of 0.02 is due to the insufficient marker information. Keeping F_{ST} =0.02 and all the other parameters but increasing the number of loci to L=100 increases the accuracy of $\widehat{K}_{PI}$ to 100%, and that of $\widehat{K}_{Ev}$ and $\widehat{K}_{Pr}$ to 72% and 82% respectively.

Unequal F_{ST} among populations does not affect the accuracy order of the three K estimators. When F_{ST} ={0.02, 0.02, 0.04, 0.04, 0.08, 0.08}, L=50 and the other parameters have the same values as in Figure 1, the accuracy is $\Pr(\widehat{K}_{Ev} = K)$=66%, $\Pr(\widehat{K}_{Pr} = K)$ =100% and $\Pr(\widehat{K}_{PI} = K)$ =100%.

Simulation 2, number of loci L

$\widehat{K}_{Ev}$ is less accurate than $\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$, especially when the number of loci L is small (Figure 2).

Simulation 3, number of populations K

$\widehat{K}_{Ev}$ cannot distinguish a panmictic population (K=1) from a structured or subdivided population (K > 1), because its statistic is undefined at K=1. When both K and L are small (Figure 3, left panel), $\widehat{K}_{Ev}$ is less accurate than the other two estimators. $\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$ have a similar accuracy when K>1. When K=1 (i.e. a single panmictic population), however, $\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$ yield the correct estimate at a frequency of about 45% and 100%, respectively. When K>10, all three estimators become highly inaccurate (Figure 3, left panel), and require more markers to yield accurate population structure inferences (Figure 3, right panel).

Simulation 4, unbalanced sampling

$\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$ are far more accurate than $\widehat{K}_{Ev}$ in the case of unbalanced sampling (Figure 4). $\widehat{K}_{Ev}$ is very vulnerable to the unevenness of sample sizes. It gives poor estimates of K even when samples from different populations are only slightly different in size.

Simulation 5, hierarchical structure

Figure 5 shows that $\widehat{K}_{Ev}$ estimates the number of archipelagos while $\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$ estimate the number of islands. It also shows that $\widehat{K}_{Pr}$ is less accurate than $\widehat{K}_{Ev}$ and $\widehat{K}_{PI}$ for different K values. Although $\widehat{K}_{Ev}$ and $\widehat{K}_{PI}$ estimate K_a and K_i respectively, they have a similar accuracy.

Simulation 6, hybridization

$\widehat{K}_{PI}$ is robust to the presence of hybrids (Figure 6). Perfect K estimates were obtained even when 32% of sampled individuals are either F1 or F2 hybrids. The other two methods also perform well, but relatively $\widehat{K}_{Ev}$ is the least accurate method even when hybrid frequencies are low. Both $\widehat{K}_{Ev}$ and $\widehat{K}_{Pr}$ show a dip in accuracy when hybrid rate is high.

Simulation 7, inbreeding

$\widehat{K}_{PI}$ is little affected by inbreeding (Figure 7). Almost perfect K estimates were obtained even when selfing rate is 80%. In contrast, $\widehat{K}_{Pr}$ works well only when inbreeding is absent or very low. Its accuracy decreases rapidly with an increasing selfing rate. It outperforms $\widehat{K}_{Ev}$ when selfing rate is negligibly small, but quickly becomes less accurate than $\widehat{K}_{Ev}$ with an increasing selfing rate. More markers do not help. Actually when selfing rate is substantial, L=20 loci leads to less accurate $\widehat{K}_{Pr}$ than L=10 (Figure 7).

***Simulation 8, many SNPs and low* F$_{ST}$**

At a low differentiation of F_{ST}=0.01, all K estimators calculated from both STRUCTURE and ADMIXTURE outputs perform poorly when the number of SNPs (L) is small (L<3200) (Figure 8). With more SNPs, $\widehat{K}_{PI}$ quickly reaches the maximal accuracy of 100%, no matter it is calculated from STRUCTURE or ADMIXTURE outputs. The other three estimators, $\widehat{K}_{Ev}$, $\widehat{K}_{Pr}$ and $\widehat{K}_{CV}$, still perform poorly even when a large number of SNPs (12800 and 204800 for STRUCTURE and ADMIXTURE, respectively) are used. The cross-validation estimator, $\widehat{K}_{CV}$, consistently yields an estimate of K=1 (i.e. no population structure), using 100-204800 SNPs.

***Simulation 9, many SNPs and variable* F$_{ST}$**

The cross-validation estimator, $\widehat{K}_{CV}$, is inaccurate when F_{ST} is small (Figure 9), as observed before (Figure 8). With $F_{ST} \leq 0.06$, the cross-validation method always yields $\widehat{K}_{CV} = 1$, much

smaller than the truth of K=5. The accuracy of $\widehat{K}_{CV}$ increases rapidly with increasing F_{ST} when it is larger than 0.06, and reaches the maximum 100% when F_{ST} =0.08. In contrast, $\widehat{K}_{PI}$ becomes perfect when $F_{ST} \geq 0.02$.

Human SSR data

$\widehat{K}_{Ev}$ consistently underestimates K, irrespective of the number of loci (Figure 10). The vast majority of estimates are $\widehat{K}_{Ev}$=2, with population Surui forming a cluster and the remaining four populations forming the other cluster. $\widehat{K}_{Pr}$ underestimates and overestimates K when L is low and high respectively, yielding a maximal accuracy (i.e. the frequency of estimates of K=5) of 60% at L=20. It is a bit bizarre that $\widehat{K}_{Pr}$ gives fewer estimates of K=5 with an increasing L when L>20. The accuracy of $\widehat{K}_{PI}$ always increases with L. The estimator is (when L>10) or is close to (when L=10) the most accurate, and yields perfect K estimates when L>20.

Discussion

In this study I proposed an *ad hoc* estimator of the number of populations represented in a sample of individuals (K), which can be calculated from the results of a STRUCTURE-like analysis. While previous estimators (e.g. Pritchard *et al*. 2000; Evanno *et al*. 2005) rely on the estimated likelihood or probability of data, the new method, in contrast, evaluates and employs the individual ancestry assignment quality as the criterion in choosing the most likely K. It is based on the observation that, in a STRUCTURE-like analysis, assuming a higher and lower than true K value leads to the splitting and merging of source populations, respectively. In both cases, the individual ancestry assignment quality is usually undermined, as characterized by inflated admixture within each replicate run and increased inconstancy across replicate runs. Loosely speaking, the parsimony index method estimates the most likely K by identifying the number of populations which yields the most consistent and the minimal average admixture.

My extensive simulations under a variety of population structure and sampling scenarios show that the new estimator ($\widehat{K}_{PI}$) outperforms the current popular estimators overall. In some difficult situations such as unbalanced sampling (Figure 4), low population differentiation (Figures 8, 9, 1), low marker information (Figure 2), hierarchical structure

(Figure 5), and inbreeding (Figure 7), the new estimator improves *K* estimation substantially. I also show that the new estimator is accurate when hybridization is present (Figure 6), and is more accurate than other estimators when hybridization is high. This seems to be a bit of surprising, given that $\widehat{K}_{PI}$ is partially based on minimizing average admixture. However, true admixture is different from false admixture. While the former is consistently inferred across different *K* values and across different replicate runs for a given *K* value, the latter is estimated only when the assumed *K* deviates from the truth and is estimated inconsistently across replicate runs. Therefore, minimizing average admixture still leads to the recovery of the actual population structure in the presence of true admixture.

My simulation confirms the conclusion that estimator $\widehat{K}_{Ev}$ is more accurate than $\widehat{K}_{Pr}$ for populations in the hierarchical island model (Evanno *et al*. 2005). While $\widehat{K}_{Ev}$ estimates the number of archipelagos (K_a), $\widehat{K}_{Pr}$ tends to estimate the number of islands (K_i). $\widehat{K}_{Ev}$ estimates K_a with an accuracy consistently and considerably higher than $\widehat{K}_{Pr}$ estimates K_i (Figure 5). The simulations also show $\widehat{K}_{Ev}$ is more accurate than $\widehat{K}_{Pr}$ for highly inbred populations (Figure 7). Unfortunately, however, the accuracy advantage of $\widehat{K}_{Ev}$ is lost in other scenarios of population structure and sampling (Figures 1~4, 6). In several realistic scenarios (e.g. unbalanced sampling, Figure 4), $\widehat{K}_{Pr}$ is much more accurate than $\widehat{K}_{Ev}$. It is unfortunate that $\widehat{K}_{Ev}$ is now widely favoured over $\widehat{K}_{Pr}$ in estimating *K*, even when there is no clear evidence that the populations are in a hierarchical structure or highly inbred.

The confusion as to which estimator gives a better *K* estimate arises because all estimators are *ad hoc* and their accuracies must be evaluated using simulated or empirical datasets. Due to the heavy computational burden, however, few studies (e.g. Evanno *et al*. 2003; Gao *et al*. 2011) were conducted to compare the accuracy of different estimators under various population structure and sampling scenarios. Typically a simulation study (like the present one) considers many different sets of parameter combinations, and simulates and analyses a large number of replicate datasets for a given parameter combination. Each simulated dataset must be analysed with different assumed *K* values, and for each assumed *K* value, a number of replicate runs (say, 20) must be conducted. The total number of STRUCTURE runs for a single dataset is $(K_H - K_L + 1) \times 20$, which is 140 when the simulated $K>3$, $K_H = K+3$, and $K_L = K-3$. If a typical run takes about 0.5 hours, this means analysing a single dataset takes about 70 hours. Analysing 50 replicate datasets simulated under a given set of parameters would take about 3500 hours. A typical figure with 8 plotting points (values

on the *x* axis) would take about 28000 hours, and the 8 figures from STRUCTURE analyses shown in this study would take about 224000 hours. It is obviously impossible to conduct a simulation study like the present one on a desktop computer. My simulation was carried out on a Linux cluster using 512 cores in parallel.

This study focussed on applying different *K* estimators to STRUCTURE (Pritchard *et al*. 2000) analyses. Other methods for population structure inference (e.g. Corander *et al*. 2003; Tang *et al.* 2005; Gao *et al.* 2007; Huelsenbeck and Andolfatto 2007; Alexander *et al.* 2009) use the same genotype data and give similar outputs such as individual ancestry proportions. For the analysis of genomic SNP data with thousands to millions of loci, STRUCTURE is too slow and much faster methods are increasingly used. Alexander *et al.* (2009) improved Tang *et al*.'s (2005) expectation maximization algorithm of a likelihood model, and implemented the algorithm in a computer program ADMIXTURE. The program runs several orders faster than STRUCTURE, and yet provides similarly good results of both ancestry assignments and *K* estimates (by the cross-validation method, $\widehat{K}_{CV}$) in some tested situations (Alexander *et al.* 2009; Alexander and Lange 2011). This study simulated genomic data and compared the performances of the cross-validation method and other methods (Figure 8). It is clear that $\widehat{K}_{CV}$ calculated from ADMIXTURE outputs behaves similarly to $\widehat{K}_{Pr}$ and $\widehat{K}_{Ev}$ calculated from STRUCTURE outputs, when populations are little differentiated and many SNPs are used. The accuracy of the three estimators is rather poor, compared with that of the parsimony estimator $\widehat{K}_{PI}$ calculated from the outputs of both programs. It seems $\widehat{K}_{CV}$ is very sensitive to the F_{ST}, and becomes accurate only for highly differentiated populations (Figure 9). At low differentiation, it is conservative and always falsely infers a single ($\widehat{K}_{CV} = 1$) panmictic population even when many markers are used (Figures 8 and 9).

The parsimony *K* estimator described in this study was implemented in a computer program, KFinder, freely downloadable from https://www.zsl.org/science/software/KFinder.

References

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), Selected papers of Hirotugu Akaike (pp. 199–213). Springer Series in Statistics. New York, NY: Springer.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655-1664.

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Beugin MP, Gayet T, Pontier D, Devillard S, Jombart T (2018) A fast likelihood solution to the genetic clustering problem. *Methods in Ecology and Evolution*, **9**, 1006-1016.

Cavanaugh JE (1999) A large- sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, **42**, 333–343.

Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

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

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS2: enhanced possibilities for the analysis of population structure. *Bioinformatics* **20**, 2363-2369.

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

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, **11**, 2571-2581.

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 Resources*, **7**, 574-578.

François O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

569 Gao X, Song P (2010) Composite likelihood Bayesian information criteria for model
570 selection in high-dimensional data. *Journal of the American Statistical Association*, **105**,
571 1531–1540.
572 Gao H, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for
573 joint inference of population structure and inbreeding rates from multilocus genotype
574 data. *Genetics*, **176**, 1635-1651.
575 Gao H, Bryc K, Bustamante CD (2011) On identifying the optimal number of population
576 clusters via the deviance information criterion. *PloS one*, **6**, e21014.
577 Guillot G, Mortier F, Estoup A (2005) GENELAND: a computer package for landscape
578 genetics. *Molecular Ecology Resources*, **5**, 712-715.
579 Hsieh HM, Huang LH, Tsai LC, Kuo YC, Meng HH, Linacre A, *et al*. (2003) Species
580 identification of rhinoceros horns using the cytochrome b gene. *Forensic Science*
581 *International*, **136**, 1–11.
582 Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure
583 with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322-
584 1332.
585 Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet
586 process prior. *Genetics*, **175**, 1787–1802.
587 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a
588 new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.
589 Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P (2002)
590 Assessing population differentiation and isolation from single nucleotide polymorphism
591 data. *Journal of the Royal Statistical Society Series B*, 64, 695–715.
592 Patterson NJ, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS*
593 *Genetics*, **2**, e190.
594 Pella J, Masuda M (2006) The gibbs and split-merger sampler for population mixture analysis
595 from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic*
596 *Sciences*, **63**, 576–596.
597 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using
598 multilocus genotype data. *Genetics*, **155**, 945–959.

599 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De
600 Bakker, PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association
601 and population-based linkage analyses. *The American Journal of Human Genetics*, **81**,
602 559-575.

603 Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of
604 population structure in large SNP data sets. *Genetics*, **197**, 573-589.

605 Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

606 Smouse PE, Waples RS, Tworek JA (1990) A genetic mixture analysis foruse with
607 incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*,
608 **47**, 620–634.

609 Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and
610 study design considerations. *Genetic Epidemiology*, **28**, 289-301.

611 Velo-Anton G, Godinho R, Ayres C, Ferrand N, Rivera AC (2007) Assignment tests applied
612 to relocate individuals of unknown origin in a threatened species, the European pond
613 turtle (Emys orbicularis). *Amphibia-Reptilia*, **28**, 475–84.

614 Wang J (2017) The computer program STRUCTURE for assigning individuals to populations:
615 easy to use but easier to misuse. *Molecular Ecology Resources*, 17, 981-990.

616 Wang S, Lewis Jr CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra
617 MV, Molina JA, Gallo C, Mazzotti G (2007) Genetic variation and population structure
618 in Native Americans. *PLoS Genetics*, **3**, e185.

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

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

622 Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323-354.

623

624 **Data Accessibility**

625 Source code (in Fortran 9x) for simulating genotype data, preparing input files for
626 STRUCTURE, running STRUCTURE, and calculating the three *K* estimators: DRYAD entry
627 DOI: ####

628 **Acknowledgements**

I am grateful to the editor, N Barton, and two anonymous reviewers for helpful comments on previous versions of this article.

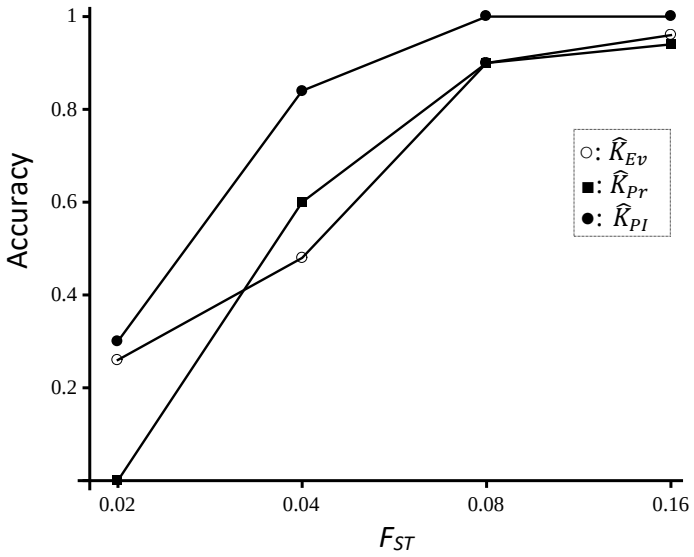


Figure 1: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of F_{ST}. A number of *K*=6 populations in the island model with F_{ST} =0.02, 0.04, 0.08 or 0.16 (*x* axis) were simulated. Thirty individuals from each population were sampled and genotyped at 20 loci, each having *A*=10 alleles.

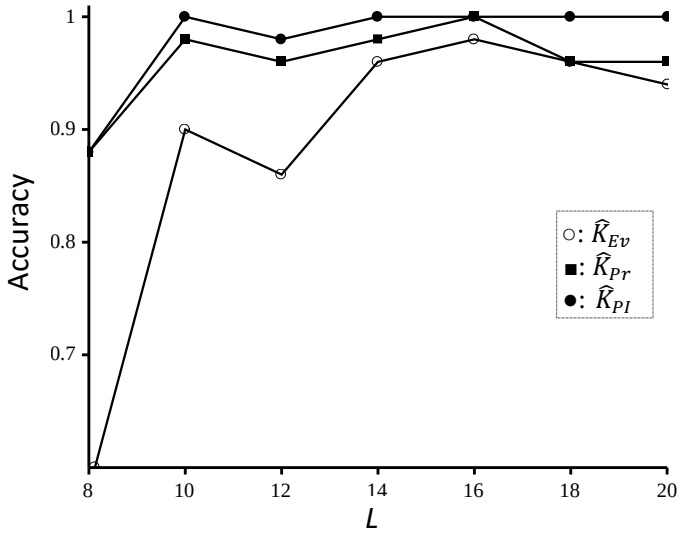


Figure 2: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of the number of loci *L*. A number of *K*=6 populations in the island model with F_{ST} =0.1 were simulated. Thirty individuals from each population were sampled and genotyped at *L*=8, 10, 12, 14, 16, 18 and 20 loci, each having *A*=10 alleles.

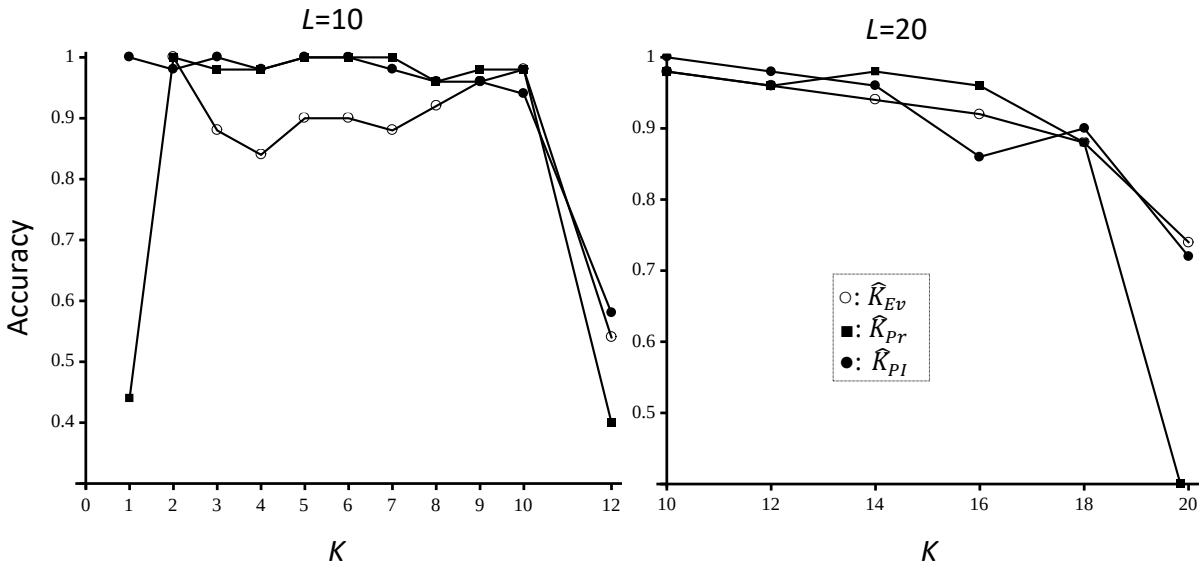


Figure 3: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of the simulated (true) number of populations (*K*). A number of *K* (*x* axis) populations in the island model was simulated, assuming F_{ST}=0.1. A number of 30 individuals were sampled from each population, and each individual was genotyped at either *L*=10 (left panel) or *L*=20 (right panel) loci, each having 10 alleles.

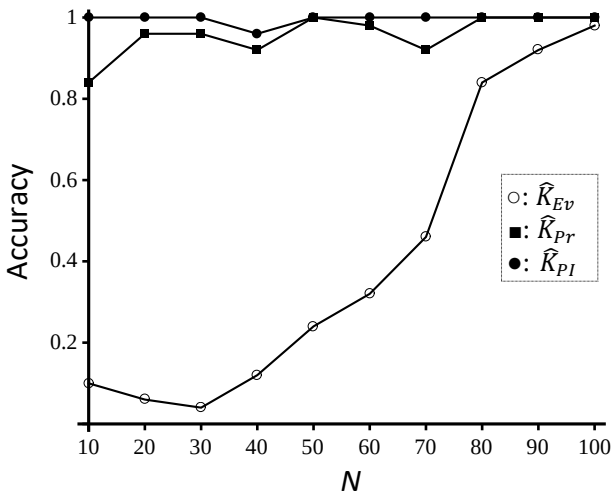


Figure 4: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of the smaller sample size *N*. Three populations in the island model was simulated, assuming F_{ST}=0.1. A number of 300-2*N*, *N*, and *N* individuals were sampled from populations 1, 2, and 3, respectively. The individuals were genotyped at *L*=20 loci, each having 10 alleles.

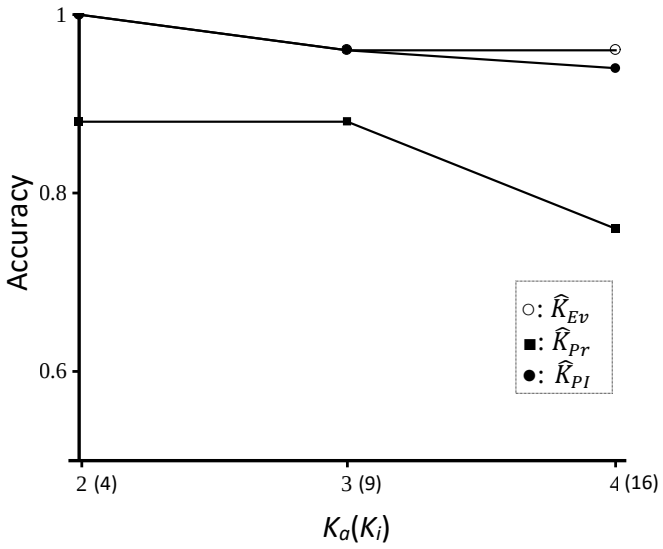


Figure 5: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of the number of archipelagos K_a or the number of islands K_i in a HI model. Note, the HI model has two true *K* values, the number of archipelagos (K_a) and the number of islands (K_i), the latter being equal to the square of the former in the simulations. Accuracy $\Pr(\widehat{K}_{Ev} = K)$, $\Pr(\widehat{K}_{Pr} = K)$ and $\Pr(\widehat{K}_{PI} = K)$ is calculated with *K* being K_a for $\widehat{K}_{Ev}$, and being K_i for $\widehat{K}_{Pr}$ and $\widehat{K}_{PI}$. The F_{ST} values among archipelagos and among islands within an archipelago are both 0.1. Thirty individuals were sampled from each island, and were genotyped at L=20 loci, each having 10 alleles.

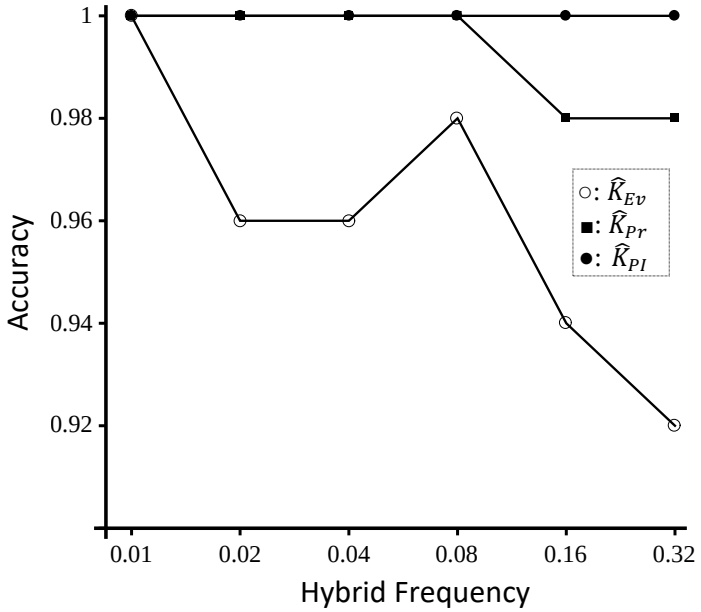
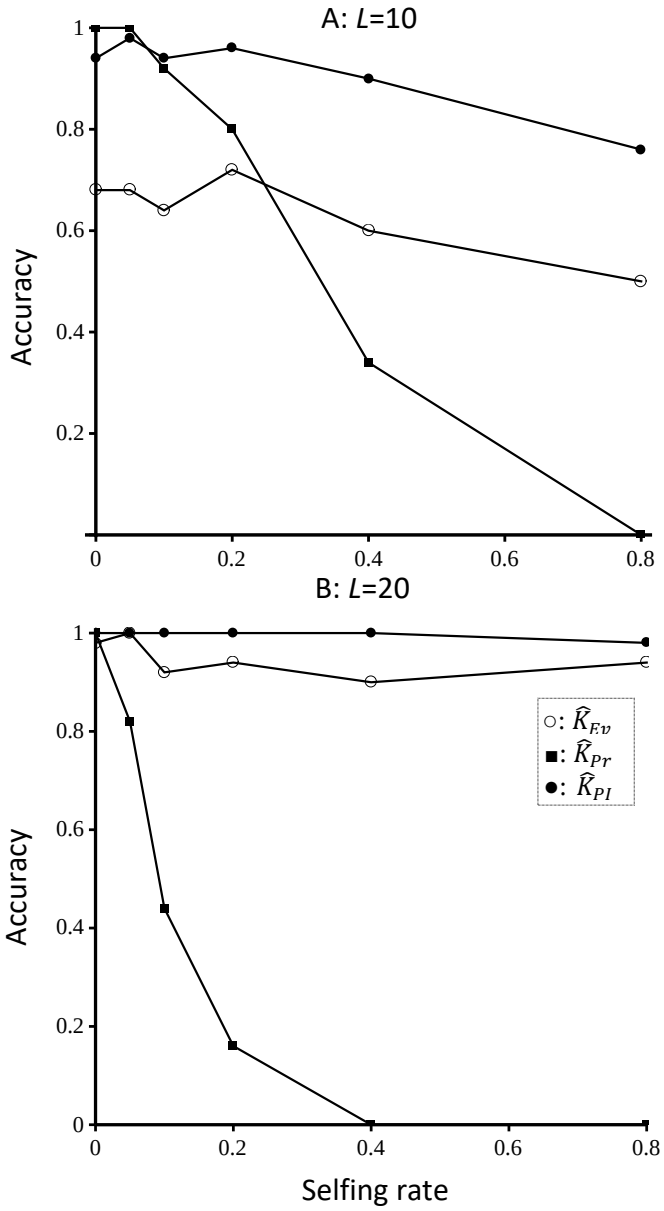


Figure 6: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a function of the frequency of hybrids (F1 and F2) in a sample of individuals. Three populations in the island model with F_{ST}=0.1 were simulated, and 30 individuals were sampled from each population and were genotyped at *L*=20 loci, each having 10 alleles.



692

693

694

695

696

697 Figure 7: Accuracy of three *K* estimators (calculated from STRUCTURE outputs) as a
698 function of the selfing rate of sampled individuals. Five populations in the island model with
699 F_{ST}=0.1 were simulated, and 30 individuals were sampled from each population and were
700 genotyped at *L*=10 (upper panel) or 20 (lower panel) loci, each having 10 alleles.

701

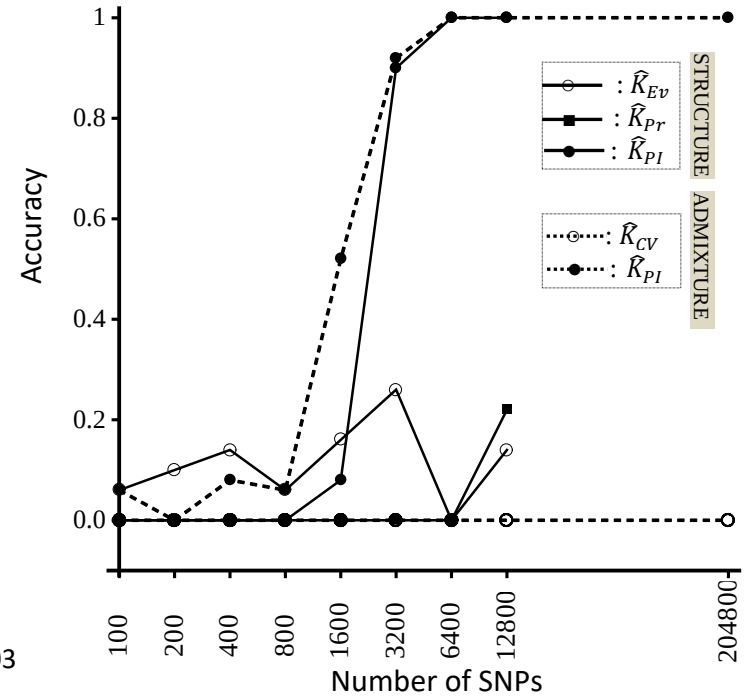


Figure 8: Accuracy of estimators $\widehat{K}_{Ev}$, $\widehat{K}_{Pr}$, $\widehat{K}_{PI}$ calculated from STRUCTURE outputs and $\widehat{K}_{CV}$ and $\widehat{K}_{PI}$ calculated from ADMIXTURE outputs as a function of the number of SNPs used in population structure analyses. Five populations in the island model with F_{ST}=0.01 were simulated, and 20 individuals were sampled from each population and were genotyped at L=100 - 204800 (x axis) loci, each having 2 alleles.

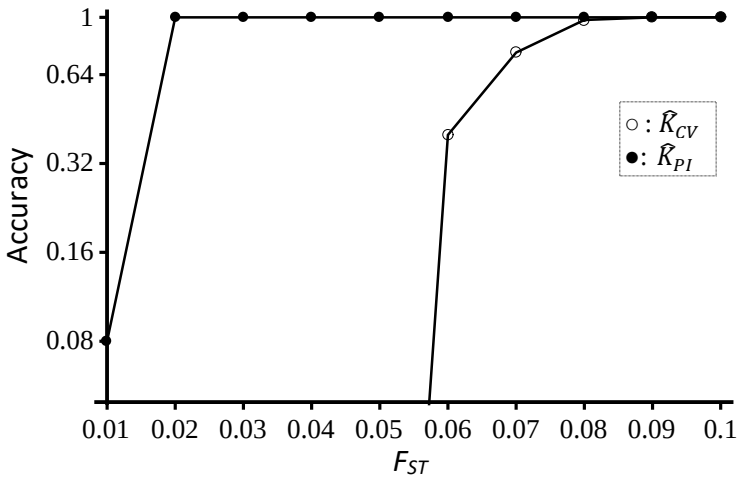
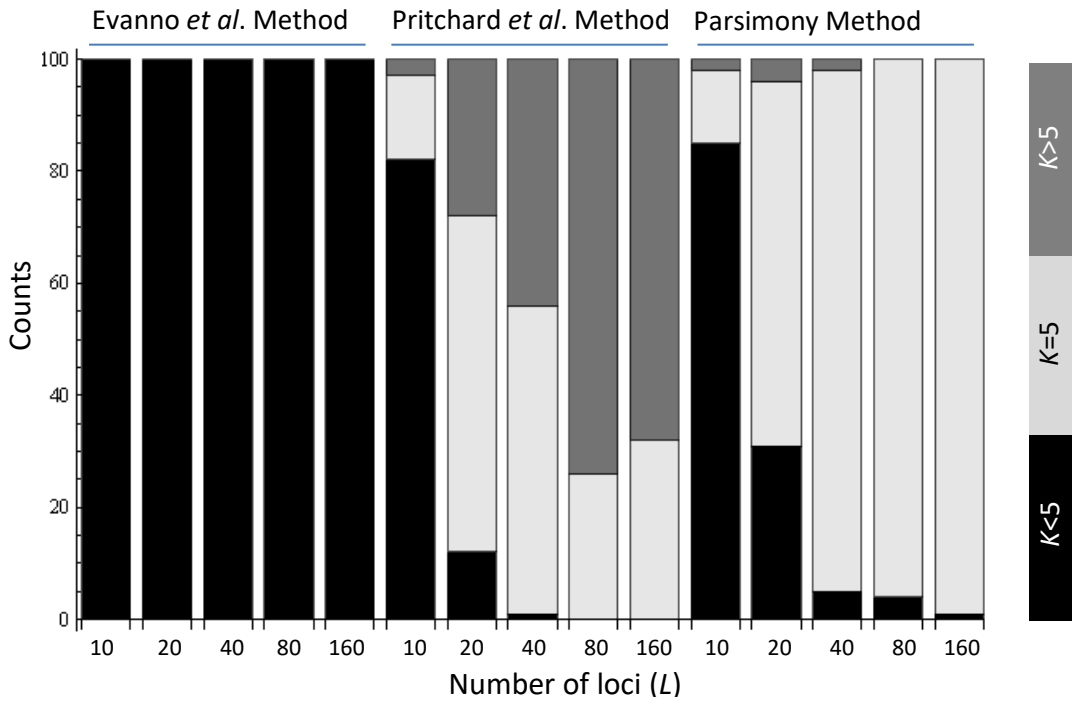


Figure 9: Accuracy of estimators $\widehat{K}_{CV}$ and $\widehat{K}_{PI}$ calculated from ADMIXTURE outputs as a function of F_{ST}. Five populations in the island model with F_{ST} varying (x axis) in the range [0.01, 0.10] were simulated, and 20 individuals were sampled from each population and were genotyped at L=1000 loci, each having 2 alleles.



722

723

724

725

726 Figure 10: Accuracy of three *K* estimators as a function of the number of loci, *L*. A number of
727 100 replicate datasets were obtained by bootstrapping (over loci) for each number of loci (*x*
728 axis) from a human SSR dataset, and were analysed by STRUCTURE for estimating *K*. The
729 original dataset has 117 individuals sampled from 5 (true *K*=5) populations.