

COMPUTER PROGRAM ARTICLE

FLOCK: a method for quick mapping of admixture without source samples

P. DUCHESNE and J. TURGEON

*Département de Biologie, Université Laval, Québec, QC, Canada, G1V 0A6***Abstract**

Identifying and estimating individual and/or population admixture is a very common objective in evolution and conservation biology. There are many situations where samples from one or many of the putatively hybridizing entities are not available or easily identified. Here we describe FLOCK, a new method especially designed to provide spatial and/or temporal admixture maps in the absence of one or several source samples. FLOCK is a non-Bayesian method and therefore differs substantially from previous clustering algorithms. Its working principle is repeated re-allocation of all collected specimens (total sample) to the k subsamples, each re-allocation being more effective than the previous one in attracting genetically similar individuals. This snowball effect, more formally referred to as a positive feedback mechanism, makes FLOCK an efficient and quick sorting process. The usage of FLOCK is illustrated with two empirical situations which have been thoroughly analysed previously with other approaches. A number of simulations were run to better assess the power of the FLOCK algorithm. Performance comparisons were made between the FLOCK and Structure algorithms. When non-negligible numbers of pure genotypes were present, the two performed equally well. However, FLOCK proved significantly more powerful in the absence of pure genotypes. Moreover, FLOCK showed more potential for fast processing. Run times were shown to increase linearly with size of total sample and with size of k , the number of reference samples from which admixture mapping is performed.

Keywords: admixture analysis, iterative method, introgression, hybridization, population re-allocation

Received 20 August 2008; revision accepted 8 January 2009

Introduction

The concept of biological population is as central to biology as it is difficult to define. On the one hand, the multiple facets of the population concept derive from the alternative cohesive mechanism, either demographic or reproductive, that is pertinent to the question one asks and the tools one utilizes (e.g. Waples & Gaggiotti 2006). On the other hand, connectivity among groups of conspecific individuals necessarily blurs the delineation of populations. From an evolutionary standpoint, spatially discrete, demographically stable and isolated populations are relatively easy to identify because they often possess distinct genetic features. However, reproductive connectivity among distinct groups of individuals is common and results in individuals having

ancestry in more than one such group. These genetic exchanges may be viewed as gene flow, introgression or hybridization, or more generally, as genetic admixture.

Identifying and estimating individual and/or population admixture is a very common objective in evolution and conservation biology. Hybrid zones are often revealed in studies of ancient or contemporary colonization patterns (e.g. Hewitt 2000) and they are widely used to study speciation mechanisms and selection processes (e.g. Barton & Hewitt 1985; Mavarez *et al.* 2006). Conservation biology also makes wide use of information about genetic admixture, for example, to evaluate the impact of supplementing wild native populations (e.g. Hansen 2002), to estimate genetic restoration potential (e.g. Hansen *et al.* 2006) or to analyse and monitor the spread of intentionally re-introduced populations or species (e.g. Hedrick & Fredrickson 2008; Jacobsen *et al.* 2008) as well as hybridizing invasive species (e.g. Boyer *et al.* 2008).

Correspondence: Julie Turgeon, Fax: 418-656-2043, E-mail: julie.turgeon@bio.ulaval.ca

Admixture analyses most often rely on the distinctive genetic characteristics of 'pure' populations or species to estimate the degree to which putatively admixed individuals resemble one or the other references (e.g. Barton & Hewitt 1985; Smouse *et al.* 1990; Dupanloup & Bertorelle 2001; Pella & Masuda 2001, 2006). However, there is a wealth of situations where samples from one or many of the putatively hybridizing entities are not available or easily identified. Indeed, pre-admixture reference samples are often missing in cases of recent admixture caused by human action. Examples include (i) hybridization of domesticated lineages with wild ancestors that produce hybrids phenotypically similar to wild specimens (reviewed in Randi 2008); (ii) invasion and asymmetrical introgressive hybridization potentially leading to the absorption and the extinction of one gene pool (Rhymer & Simberloff 1996; Perry *et al.* 2001); (iii) admixture of recently diverged species following environmental homogenization (reviewed in Seehausen *et al.* 2008); (iv) mixing of regional cultivars or domestic races aided by recent commercial exchanges among previously isolated regions (Allinne *et al.* 2008); and (v) intentional supplementation of declining wild populations without historical tissue collection of the indigenous populations (e.g. Taylor *et al.* 2007).

Here we describe FLOCK, a new method especially designed to provide spatial and/or temporal admixture maps in the absence of one or several sources, sometimes called 'pure', samples. Specifically, its primary goal is to capture the remnant signal of past, pre-admixture, differentiation among formerly genetically distinct groups. FLOCK attempts to do so by grouping contemporary admixed specimens along their ancestral differentiation lines. Once performed with reasonable resolution, this reconstruction can then be used to estimate levels of individual and sample admixture as well as to draw spatial and temporal admixture maps. It should be noted that the expression 'without source samples' does not mean that FLOCK can be used to identify nonsampled sources. FLOCK will deal with samples comprising specimens of various degrees of admixture but nonsampled genetic components will be ignored. Loosely speaking, it is a 'clustering' method in that it starts out with all collected specimens and proceeds to part this one global sample (S) into some number k of genetically differentiated subsamples. However, its working principle does not involve a probabilistic walk through the space of all possible k -partitions of sample S as is the case with clustering algorithms such as Structure (Pritchard *et al.* 2000), BAPS (Corander *et al.* 2003) and NewHybrids (Anderson & Thompson 2002). Partition (Dawson & Belkhir 2001) is another clustering program. However, since it is assuming Hardy-Weinberg and linkage equilibrium, it is not suitable for the analysis of samples originating from genetic groups in the process of becoming admixed and, consequently, that are in nonequilibrium, dynamically transient, states.

Generally speaking, FLOCK is a non-Bayesian method and therefore differs substantially from previous clustering algorithms. Its working principle is repeated re-allocation of S to the k subsamples. Since each subsample will tend to attract similar individuals, repeated re-allocation will exert a filtering effect by progressively building up homogeneity within and differentiation between subsamples. As the subsamples get more and more differentiated, they become more and more effective attractors of whatever specimens are still in the 'wrong basket' (see Fig. 1). This snowball effect, more formally referred to as a positive feedback mechanism, makes FLOCK an efficient and quick sorting process. More intuitively, it may be said that FLOCK does not divide up sample S but sets the stage for specimens to agglutinate according to resemblance as in 'birds of a feather flock together'.

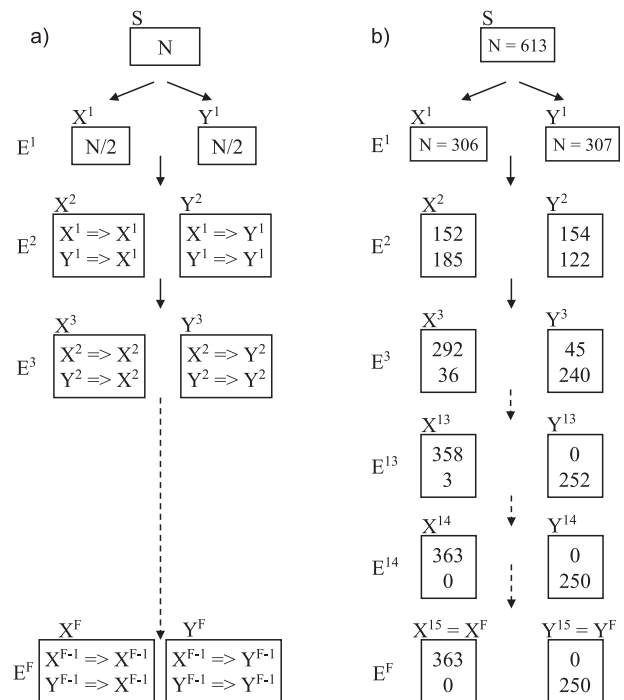


Fig. 1 Illustration of the FLOCK procedure. (a) The initial sample S is randomly divided, here into $k=2$ subsamples to form X^1 and Y^1 . Iterative re-allocation (\Rightarrow) generates successive states (E^i) with increasingly different subsamples X^i and Y^i . The final X^F and Y^F subsamples may serve as reference groups to allocate individuals from S and thereby estimate and map individual or group admixture levels. (b) the re-allocation matrices obtained with the Lake ciscoe data from Turgeon & Bernatchez (2001b). The numbers result from applying the re-allocation procedures described formally in a). For instance, in E^2 , 152 stands for the number of X^1 genotypes re-allocated to X^1 ($X^1 \Rightarrow X^1$) while 185 stands for the number of Y^1 genotypes now allocated to X^1 ($Y^1 \Rightarrow X^1$). The total number of X^2 is therefore 337.

FLOCK algorithm and rationale

Let us suppose that some mixed sample S of genotypes is made up of specimens variously admixed between two originally distinct genetic groups, say G_1 and G_2 . Here 'distinct' means that those groups are genetically differentiated. The purpose of the FLOCK procedure is to separate the specimens that are most alike G_1 from those that are most alike G_2 . Moreover, we assume that, save the genotypes, no information is available that could inform the separation process.

Suppose the mixed sample S is divided into two arbitrary subsamples of equal sizes, X^1 and Y^1 , and that the genotypes of X^1 and Y^1 are re-allocated, resulting in the production of subsamples X^2 and Y^2 . Re-allocations are performed following multilocus maximum likelihood (Paetkau *et al.* 1995) and the leave-one-out procedure. Now there is a very high probability that either X^1 or Y^1 is more similar to G_1 than it is to G_2 . Without loss of generality, assume X^1 closer to G_1 . Then any G_1 -like specimen will be more likely re-allocated to X^1 than to Y^1 . Therefore, the genetic contents of X^2 relative to the contents of X^1 will be more G_1 -like. Clearly one could focus on G_2 -like specimens and apply an analogous line of reasoning.

Through repeating the above procedure, one would obtain sequences of X^i and Y^i , one of which increasingly more G_1 -like and the other increasingly more G_2 -like (see Fig. 1). This iterative process starting from a random division of the sample S is the FLOCK procedure. However, FLOCK is not restricted to separation into two reference groups. It can in fact be applied to any number k of component groups within the mixed sample S .

Re-allocation number matrices are a handy representation of the output of FLOCK as it performs re-allocations in succession (Fig. 1). They are especially useful in assessing output stability and therefore the number of re-allocations to reach it. Re-allocation number matrices, stability criteria and sufficient number of re-allocations are described in more detail in the Appendix. Another useful representation of FLOCK results is the difference in allocation log-likelihoods to the reference groups of each individual (LLOD score). The representation of individual or sample mean LLOD scores over space, time, or any other meaningful ordination axis, generates a map of admixture level in S .

Validation of k groups

When FLOCK is run with some k value, it necessarily separates the mixed sample S into k groups. Some of these groups may not reflect existing biological structure but some chance structure at the sample level. There is therefore a need to validate the k groups.

Validation can be performed in several ways. Each validation involves information other than the one provided by the mixed sample of genotypes (S). Two types of validation are briefly discussed below:

Allocating over several samples and testing for random composition

Assume the mixed sample S is made up of a collection of several empirical samples. The empirical samples may cover some geographical area or temporal period. In order to validate the k reference groups, simply allocate the genotype samples to the k groups ($G_1, G_2, G_3 \dots G_k$) and compile the number of specimens within each sample which are allocated to each reference group, thus building an $r \times c$ contingency table where r is the number of reference groups and c is the number of empirical samples. Then test for random allocation to the reference groups across empirical samples using a traditional chi-squared (χ^2) test (Sokal & Rohlf 1981). In an extensive simulation experiment, this test has been shown by Ryman & Jorde (2001) to combine an α error keeping close to the intended one and high power when compared to several other approaches including Fisher's exact test. Traditional chi-squared (χ^2) testing may be easily performed within widely used computer packages such as Excel. As is customary, P values below 0.05 are taken to be an indication that reference group composition is unlikely to be random across empirical samples and therefore validates the k groups. To illustrate this procedure, here is the allocation composition of the Lake ciscoe samples (Fig. 1, and see below under *Study cases*) to the two reference groups:

| | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|
| 26 | 20 | 20 | 25 | 20 | 30 | 20 | 48 | 20 | 20 | 17 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 47 | 28 | 41 | 3 | 3 | 1 | 3 | 1 | 0 | 0 | 2 |
| 1 | 1 | 4 | 22 | 17 | 19 | 25 | 49 | 30 | 28 | 18 |

The P value obtained from chi-squared testing was $< 10^{-10}$.

When $k > 2$ and a significant P value is obtained, chi-squared validation should also be performed on each pair of reference groups to also test for pairwise distinctness.

Comparing phenotypic traits

Compute the average values of chosen phenotypic trait(s) for reference groups formed by FLOCK and test their differences for statistical significance. In most instances, standard Student t -tests or F tests will be appropriate. Note that contrasting traits are more likely to stand out and be more easily identified after grouping of specimens.

Some examples — study cases

We now illustrate with two empirical examples the use of FLOCK. The first example relates to a natural situation

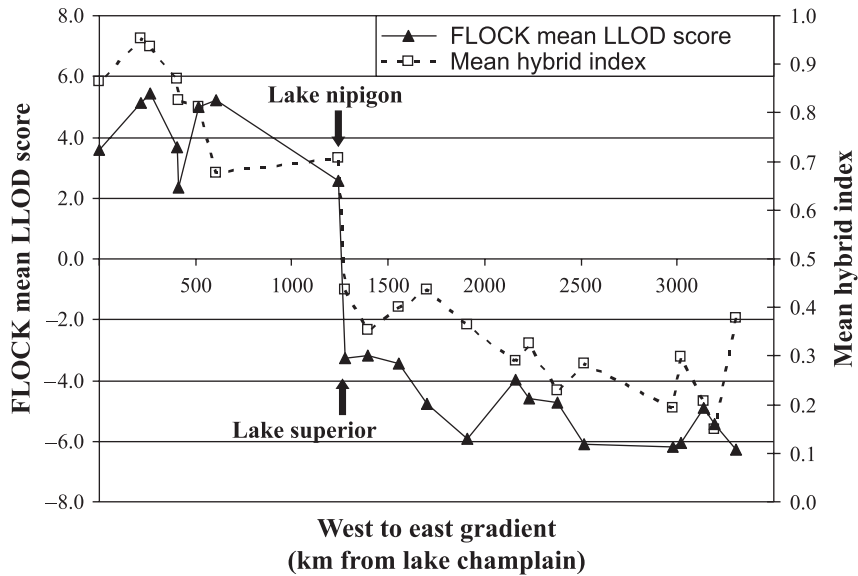


Fig. 2 Admixture levels between glacial races of Lake ciscoes (Turgeon & Bernatchez 2001b) along a series of geographically ordered sites. The grey curve reports the original admixture estimation while the black line traces the mean log-likelihood difference between the two reference groups obtained by FLOCK. The original hybrid index (Turgeon & Bernatchez 2001b) was calculated by assigning a score of one/zero for the presence of an eastern/western allele at loci exhibiting clinal patterns of frequency along the east-west axis gradient.

where the true history of divergence and admixture is unknown, while the second case refers to the detection of recent anthropogenic population admixture in the context of population supplementation. In both cases, FLOCK is used to obtain reference groups within an introgression context and in the absence of all or part of the source samples. The reference groups are subsequently used to either map admixture geographically or to estimate the stocking/nonstocking components at site and individual levels. All runs reported in this article were performed in Maple version 9.5 (MapleSoft 1981–2004) on an Intel(R), 2.66 GHz, 2.99 Go CPU.

Historical introgression between Lake ciscoes (Coregonus artedii) glacial races

In the Holarctic, pleistocene glaciations forced southward population shifts into refugial ice-free areas. As a result of allopatric isolation in distinct refuges, species lineages developed into genetically differentiated glacial races (Hewitt 2000). These races later established their current geographical distribution and intermixed at various degrees contingent upon the timing and availability of recolonization routes. In North America, proglacial lakes formed at the margin of the receding Laurentian ice sheet c. 6000–8000 years ago and provided formidable postglacial dispersion avenues for fish and other aquatic biota. There is compelling evidence that freshwater fish races developed in several refugia and that they dispersed and intermixed during the recolonization process (reviewed in Bernatchez & Wilson 1998).

As per many other fish species, analysis of mtDNA polymorphism in the Lake ciscoe, *Coregonus artedii*, suggested the existence of an Atlantic and a Mississippian glacial

racess. However, races co-occurred throughout most of the range, and there were no sites where the presence of a single race could undoubtedly be assumed (Turgeon & Bernatchez 2001a). Likewise, microsatellite polymorphism distribution suggested widespread introgression between the Atlantic and Mississippian races, but the analyses were hindered by the lack of 'pure' reference samples (Turgeon & Bernatchez 2001b). Nevertheless, analyses relying on bimodal allele size distributions at multiple loci provided evidence for a continental-wide cline of admixture levels decreasing from west to east, thus supporting the admixture hypothesis. Importantly, several analyses suggested a clinal break indicating that Lake Nipigon and Lake Superior, which are currently interconnected, were dominated by fish of Mississippian and Atlantic origin, respectively (Turgeon & Bernatchez 2001b, 2003).

The FLOCK procedure has been applied with $k = 2$ to the same genetic data set, including all sampled specimens from all lakes, and stable reference groups were obtained after 15 iterations (see Fig. 1b). Total run-time was 15.2 s. A k validation was performed based on the allocation number distribution over samples. The resulting chi-squared P value was $< 10^{-10}$. The allocation numbers (Fig. 1b) and the per-lake average LLOD scores were then calculated (Fig. 2). The resulting admixture mapping was in total agreement with the previous analyses in revealing a decreasing eastward contribution of one reference group, the Atlantic glacial race, as well as a clear clinal break in the vicinity of Lake Nipigon and Lake Superior. Besides its simplicity and speed of execution, an important feature of FLOCK is that the results relied on the whole data set rather than on a few loci which happened to carry the right information. Had these loci with bimodal allele size not been scored, the investigators might have overlooked their clinal structure.

Evidence of population supplementation in grayling (Thymallus thymallus) of Lake Saimaa (Finland)

A common conservation or management strategy to assist declining species is to supplement local populations with captive-bred individuals. Most often, the parents of these individuals are not from the local population, such that their offspring can bring foreign genes into the supplemented population. As such, the mere presence of foreign alleles in a supplemented population, as well as the degree of individual admixture, provide a way to assess the success of the supplementation procedure.

In Finland, fish from a major hatchery on Lake Saimaa came from the Puruvesi region of the lake. The Puruvesi hatchery was used to supplement the Pielinen, Etelä-Saimaa and Höytiäinen populations of Lake Saimaa since 1986. For two such sites (Pielinen and Etelä-Saimaa), historical fish scale samples from the pre-supplementation period were available as reference for the local population gene pools. Moreover, a temporal series of samples from the hatchery broodstock used to produce the supplemented individuals was available. Using these reference samples, Koskinen *et al.* (2002) performed allocation and exclusion analyses to classify contemporary individuals from these supplemented sites as representative of either indigenous or hatchery populations. They also estimated the individual degree of admixture between contemporary (post-stocking) indigenous fish and the hatchery broodstock by performing Bayesian analyses using the Structure software. Results indicated that local populations retained their genetic distinctiveness and were not much introgressed with the stocked fish. Nevertheless, in Pielinen, there was also evidence that introgression was increasing with time. However, results from Etelä-Saimaa were not conclusive because of the very large confidence interval around admixture level, even in pre-stocking historical samples.

We performed an analysis with FLOCK without using the historical reference samples. Namely, we assumed that neither the hatchery nor the Etelä-Saimaa and Pielinen historical source samples were available. The procedure was run independently with $k = 2$ for each of the three local stocking situations. In each case, the global sample S was a combination of either site Etelä-Saimaa, site Pielinen or site Höytiäinen specimens with local specimens from Puruvesi (and not from the hatchery itself). For Pielinen, specimens were divided according to the sampling dates of 1998 and 2001 and FLOCK was run independently for each date.

For each local stocking situation (Etelä-Saimaa + Puruvesi, Pielinen + Puruvesi, Höytiäinen + Puruvesi), two genetic groups were obtained by applying FLOCK: Ref_Puruvesi, comprising most Puruvesi specimens, and Ref_LocalSite. Note that these 'reference' groups should not be confused with empirical, 'pure', reference samples which were not considered in the analysis. The likelihood log differences

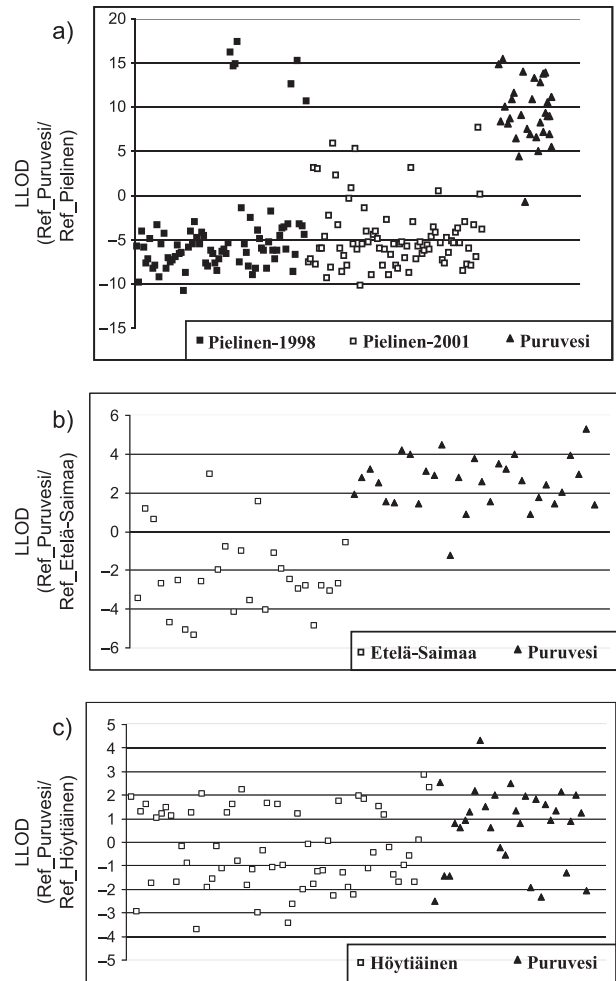


Fig. 3 Admixture analysis of European graylings in Lake Saimaa, Finland, showing LLOD scores for specimens from Puruvesi and specimens from local origin. (a) Puruvesi + Pielinen; (b) Puruvesi + Etelä-Saimaa; (c) Puruvesi + Höytiäinen. LLOD scores are shown for Puruvesi vs. local sources. Genotypes from Koskinen *et al.* (2002). Note that the x-axis does not refer to a measuring variable but only serves to spread out LLOD distributions of individual genotypes and to separate blocks of specimens on the basis of sample location/date.

(LLOD scores) between Ref_Puruvesi and Ref_LocalSite were subsequently calculated for each specimen. These results are shown in Fig. 3 while Table 1 shows the allocation matrix for each of the three stocked sites. The most likely introgressed specimens are those having LLOD values within the main range of the Puruvesi LLOD values. Within each of the Pielinen and Etelä-Saimaa situations, all but a few of the Puruvesi specimens were allocated to a single reference group (Ref. 2 in Table 1). On the other hand, most of the local specimens were part of the other reference group, Ref. 1, but some belonged to Ref. 2 (Table 1), as expected within a stocking context. The LLOD map for

Table 1 Number of European graylings assigned to and excluded from reference populations in Koskinen *et al.* (2002) or allocated to reference groups by the FLOCK method. See text for details

| Lake sector | Koskinen <i>et al.</i> (2002)* | | | FLOCK | | |
|--------------|--------------------------------|--------------|--------------|------------|--------|--------|
| | Source | Local | Hatchery | Source | Ref. 1 | Ref. 2 |
| Pielinen | Local 1998 | not reported | not reported | Local 1998 | 70 | 7 |
| | Local 2001 | 58 | 3 | Local 2001 | 69 | 10 |
| | Hatchery† | 0 | 76 | Puruvesi† | 1 | 30 |
| Etelä-Saimaa | Local | 13 | 4 | Local | 23 | 4 |
| | Hatchery | 0 | 49 | Puruvesi† | 1 | 30 |
| Höytiäinen | Local | not analysed | not reported | Local | 36 | 24 |
| | Hatchery | | | Puruvesi† | 9 | 22 |

*In Koskinen *et al.* (2002), specimens had to be assigned to contemporary samples of the local population and excluded from reference hatchery samples. Totals differ from the FLOCK results because many specimens did not comply with the requirements of being assigned and excluded. †Koskinen *et al.* (2002) used a temporal series of the Puruvesi hatchery samples as reference while FLOCK used wild local samples from near the hatchery.

Pielinen shows an evolution from a small number of 1998 Pielinen individuals that are very much of type Puruvesi towards a larger number of 2001 specimens that are clearly introgressed but with lower LLOD scores (Fig. 3a). This result is consistent with the earlier findings using source samples and also with the evolution of introgression under this supplementation regime, that is a progressive dilution of the genetic contribution from the Puruvesi stock. In Etelä-Saimaa (Fig. 3b), four local fish were allocated to the Puruvesi reference group, suggesting introgression with Puruvesi genes, while all others were allocated to the other reference group (Table 1, Fig. 3b). These results paralleled those of Koskinen *et al.* (2002) but offer a more global portrait of introgression in Etelä-Saimaa. Indeed, these authors had not been able to interpret admixture results because of large confidence intervals around q -value estimates by Structure, even when using the historical reference samples. The results for Höytiäinen (Fig. 3c) were more tangled up, the two reference groups attracting specimens from Puruvesi. Therefore, it seems that either introgression was more complete in Höytiäinen or that the Höytiäinen and Puruvesi sites were genetically much similar from the start. However, the introgression map for the Höytiäinen specimens could still be used to choose the purest candidates for supplementation purposes.

The impact of total sample size N on execution time and on precision

In order to explore the connection between N , the size of S , and run time with FLOCK, the ciscoes full data set ($N = 613$) as well as reduced sets of $N = 30, 40, 50, 75, 150, 300$ were run until stability was reached. Note that each reduced sample has been selected at random. Run times and number of iterations were recorded for each trial

(Figure S1, Supporting information). Also, admixture mappings were drawn for sample sizes $N = 50, 75, 150, 300$ and 613 (Fig. 4a).

As expected, the run times were a (quasi) linear function of sample size. This comes largely as a result of re-allocation duration being directly proportional to the number of re-allocated individuals. The connection between sample size and the number of iterations to reach stability was not clear but this seemed not to substantially modify the linearity of the sample size vs. execution time relationship. As for the admixture mappings, they were quite similar across the different values of N and they basically conveyed the same clinal introgression structure for Lake ciscoes. The relationship between number of reference groups (k) and execution time has not been explored since its linearity appears trivial given that re-allocation with k groups consists for the most part in calculating the likelihood of each genotype within each of the k subgroups.

The impact of genetic information contents

Any type of genetic marker, whether dominant or codominant, can be used to run the FLOCK procedure. The relationships between parameters defining the level of genetic information such as number of loci, number of alleles per locus, allelic frequency distribution and the efficiency of the FLOCK procedure are clearly the same as those pertaining to simple allocation. An extensive description of those relationships can be found in Bernatchez & Duchesne (2000).

With admixture mappings, lowering genetic information level will tend to reduce the definition of the mapping. This, of course, might cause important patterns to be flattened out to a point where they can be no longer distinguished.

Wishing to gain some insight into the consequences of reducing genetic information, we ran FLOCK with several

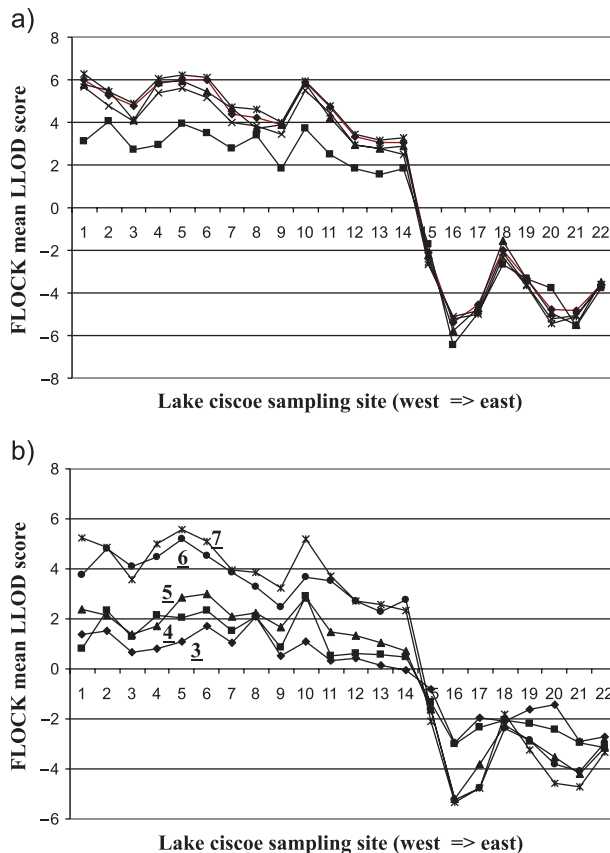


Fig. 4 Mean LLOD scores in samples of Lake ciscoes ordered from west to east for (a) randomly chosen ciscoe samples of sizes $N = 50$, 75, 150, 300 and 613. The lowest curve corresponds to $N = 50$; and (b) randomly chosen sets of loci (3, 4, 5, 6 and full data set of 7 loci) and randomly chosen 100 specimens as starting mixed sample for each set of loci (underlined number near curves indicate the number of loci).

randomly chosen sets of loci taken from the original ciscoes data set. Numbers of loci were 3, 4, 5, 6 and 7 (full data set) and 100 specimens were randomly chosen as a starting mixed sample S for each set of loci. Admixture mappings were computed and plotted for each set (Fig. 4b).

Although the mappings with the lower numbers of loci were, as predicted, flatter than the ones with more loci, they still retained the basic west–east introgression axis and even the previously documented Lake Nipigon–Lake Superior clinal break. Given that the poorest set contained only three loci and therefore very limited allocation power, this may look surprising at first sight. However, contrary to individual allocation, introgression mapping rests on the totality of specimens and, in addition, integrates the relationship between genotype samples and their location thus adding a considerable amount of information. Finally, global patterns may remain discernible even after several substantial local changes resulting from a loss of resolution.

Simulations

To better assess the power of the FLOCK algorithm given various levels of available genetic information and various degrees of admixture, a number of simulations were run. Also, performance comparisons were made with the Structure software based on an identical set of parameter values. Structure was chosen because its use is currently very widespread. Indeed, it has become a standard tool to analyse genetic structure and most investigators would use it to try to unravel the structure of admixed samples. Moreover, Structure explicitly proposes an admixture parameter with two possible values: Default and No Admixture.

Admixture building devise

First, two sets of allelic frequencies were randomly generated for 14 loci. The number of alleles for each locus were, respectively, 12, 7, 7, 7, 16, 10, 10, 7, 14, 10, 10, 7, 12, 14. This set of number of alleles spans a frequently encountered range found in genetic structure studies based on microsatellites. A table of allelic frequencies for each locus and each set is provided in Table S1 (Supporting information). Then two (pure) groups, say A and B, of 500 artificial genotypes were randomly generated, each from a different set of allelic frequencies. These genotypes made up generation 0 (G_0) and, of course, there were no admixed genotypes at this stage.

To build G_1 , 50 genotypes were chosen at random from each of A and B and put together to make up a ‘hybrid zone’ of 100 specimens (hereafter called H specimens). Then the 450 pure specimens of the A zone were made to breed together in the following way: two genotypes were chosen at random and each provided one randomly chosen allele for each locus to produce one offspring. This was carried out 450 times to obtain 450 pure A descendants. The same breeding scheme was used with the 450 specimens of the pure B zone and with the 100 specimens of the hybrid zone (H). Therefore, the hybrid zone of G_1 comprised an expected proportion of 0.5 F_1 hybrids, 0.25 pure A and 0.25 pure B genotypes and the expected proportion of hybridized specimens among all 1000 G_1 specimens was $0.5 \times 100/1000 = 0.05$.

To produce G_2 , 50 pure A and 50 pure B specimens (from G_1) were added to the hybrid zone of G_1 and then breeding took place within each zone (400 pure A, 400 pure B and 200 H specimens). The resulting hybrid zone of G_2 contained F_1 , backcrosses, pure A and pure B specimens.

G_3 and all subsequent generations were produced based on the previous generation following the same recurrent procedure. The expected proportion of hybridized genotypes as a function of generation index is shown in Fig. S2 (Supporting information) and calculation details are provided

as Appendix S1 and Table S2 (Supporting information). Note that both the proportion of hybrid individuals and their average level of admixture increase with each generation. In the long run, almost all specimens will show equal A and B ancestries. In fact, admixture within these simulations increases much more quickly, generation wise, than it does in the vast majority of natural settings. However, the admixture dynamics remains essentially the same although, for obvious practical reasons, it is greatly accelerated.

G10 marks a turning point in the admixture building process since it is bred in a panmictic mode from 1000 genotypes, all in the H zone, and among which stand very few pure A or B genotypes. Thus from G10 on, the smearing process gives way to a homogenization leading to a gradual loss of genotypes with skewed ancestry or, in other words, tending towards 50–50 ancestry. Consequently, the ancestral differentiation signal should be expected to decay quickly after G9 thereby severely limiting the power of tools designed to recover some of the pre-admixture structure. This is why the performance comparisons between FLOCK and Structure bear on generations 9–10–11. Note that G9, the last fully smeared generation, still comprises about 11% of pure A or B genotypes while G10 and G11 hold only about 0.5% and 0% of purebreds, respectively. As long as some purebreds are available, the ancestral signal is kept intact and will generally be easily captured. Therefore, it is the absence or quasi-absence of purebreds that represents the real challenge for admixture mapping algorithms.

Performance assessment

Three sequences of 11 generations (1–11) were produced, one with a choice of six loci, one with 10 loci and one with the full set of 14 loci, each sequence generated anew from the same G0. At G0, the level of differentiation between A and B, as measured by F_{ST} , was 0.070 (Belkhir *et al.* 2004). FLOCK was run assuming $k = 2$ on all 12 generations (0–11) of the six loci sequence and on generations 9–10–11 of the 10 and 14 loci sequences. Structure was run assuming $k = 2$ on generations 9–10–11 of each of the three sequences with ‘50 000 burn-in period and 100 000 reps’. All other parameters and priors were set to default. All MCMC chains converged properly. As for FLOCK, no values other than k are requested.

Performance of the FLOCK algorithm was assessed by allocating the pure genotypes of G0 to the two reference sets produced from 30 re-allocation matrices. The output variable was the sum of the largest number of the original G0 genotypes allocated to each reference set divided by the total number of genotypes allocated (1000). This proportion, say P_{anc} , is a simple, straightforward, way of measuring to what extent the differentiation between FLOCK reference

sets followed the ancestral, pre-admixture, differentiation. The same output variable was used to assess the performance of the Structure program. However, since Structure does not provide separate reference groups explicitly, the latter were built from the q -values simply by assigning the genotypes with $q < 0.5$ to one group and those with $q > 0.5$ to the alternate group. This dividing line is very similar to the one used by FLOCK since it sends each genotype to a reference group as soon as it shows a higher likelihood in that group (LLOD threshold = 0).

Results

First, it must be noted that classic, *a la Paetkau*, re-allocation based on known source samples of the pure genotypes of G0 was perfectly accurate even with six loci. Therefore, the pre-admixture A vs. B dividing line is initially crisp but becomes fuzzier and fuzzier as the admixture process goes on. The FLOCK results for the complete six loci generation sequence are shown in Fig. S3 (Supporting information) where P_{anc} is mapped as a function of the generation index. P_{anc} scores for generations 0, 1, 2 ... 8 are all above 0.97. P_{anc} decreases slightly to 0.941 with G9 despite a low proportion (5%) of each type of pure genotypes. As expected from the previous analysis of the admixture process, the P_{anc} score undergoes a significant drop at G10 (0.870) and a dramatic one at G11 (0.644).

Hereafter, only the Structure results with admixture parameter set to its *Default* value, as opposed to *No Admixture*, will be discussed since only small and nonsystematic performance differences between the two prior admixture settings were observed. Moreover, one would expect most potential users to choose the *Default* value within an admixture analysis context. The results for the G9–10–11 sequences (see Fig. 5) show that both FLOCK and Structure performed very well at G9 even with 6 loci and despite extensive admixture (89%). At G10, the FLOCK performance as measured by P_{anc} did not improve with number of loci and remained stable within the 0.86–0.88 range. By contrast, the P_{anc} score for Structure jumped from 0.648 at six loci to reach the same performance level as FLOCK when provided with 10 and 14 loci. At G11, after two episodes of genotype homogenization, FLOCK scored 0.644 with six loci but then improved slightly to reach 0.690 with 14 loci. Structure also improved with 14 loci but always scored lower on P_{anc} than FLOCK although the differences remained within the 0–0.100 range.

Average execution run times were 12 min for Structure and 4 min for FLOCK. However, FLOCK is currently coded in the programming language Maple and so runs much slower than it would if coded in a compiled programming language, for example, Java, C, C++ or Fortran. Seeking to estimate the execution time ratio between MAPLE and C++, we ran a parental allocation with PAPA (written in

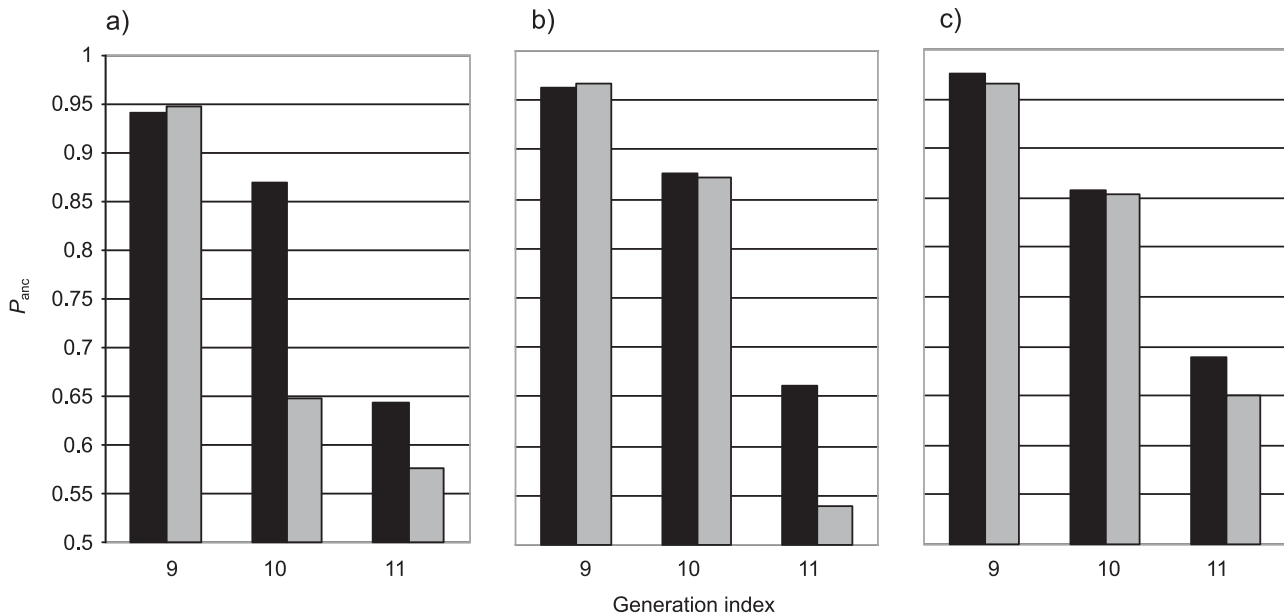


Fig. 5 P_{anc} as a function of generation index for generations 9, 10 and 11. P_{anc} scores for FLOCK and Structure are represented in black and grey, respectively, (a) with six loci (b) with 10 loci (c) with 14 loci.

C++; Duchesne *et al.* 2002) and with an equivalent Maple code. The Maple code took approximately 1200 times longer to perform the same allocation task. Therefore a rough estimate of the run time ratios between FLOCK and Structure is $12/4 \times 1200/1 = 3600/1$ and so a conservative prediction would be that FLOCK should run at least 1000 faster than Structure once coded in a compiled programming language.

Interpretation

Dividing the G9 genotypes along pre-admixing lines turns out to be an easy task for either algorithm. Apparently, this is due to the original generation (G0) being clearly differentiated even with six loci and also to the presence of 50 pure A and 50 pure B genotypes greatly helping both algorithms to pick up a clear pre-mixing differentiation signal. However, G10 with its expected number of three purebreds of each type is a more difficult case and it looks as if 0.88 were an upper bound for P_{anc} . However, this bound was reached with a lesser number of loci (six) with FLOCK than it did with Structure. By G11, the ancestral signal is much weaker and it is doubtful that adding extra loci would bring P_{anc} scores significantly above 0.70. However, the performances of both FLOCK and Structure did improve slightly with number of loci, with FLOCK always doing a little better.

In practice, extremely admixed populations are considered of a single component. Population structure analysis will be most difficult with admixture levels standing in the intermediate range, when the full smear of admixture is

about to be lost. The performance differences between FLOCK and Structure are most important when the ancestral differentiation signal is of intermediate level.

In a nutshell, the simulations have shown that given a sizeable number of A and B purebreds, recovering the ancestral differentiation signal was equally easy for FLOCK and Structure. However, FLOCK proved significantly more powerful when pure genotypes were few or absent. Moreover, the potential for fast processing definitely stood on the FLOCK side.

Summary

Situations where two or more genetically distinct genetic entities have undergone introgression over time, such that pure samples may no longer be available, are frequently encountered. FLOCK is an algorithm especially designed to provide spatial and/or temporal admixture maps in the absence of one or several source samples. First, it builds k reference samples from the totality of sampled specimens. The reference samples are obtained by iterative re-allocation starting with a random division of all specimens into k subsamples. The reference samples are subsequently used to compute and map log likelihood scores onto geographical or chronological domains. Reference samples have to be validated through statistical testing involving information other than the mixed sample of genotypes (S).

FLOCK proved an efficient, rapid method. FLOCK noticeably refined some published results without using pure samples. Moreover, the number of loci needed to reveal general admixture patterns turned out to be surprising

low. Simulations showed FLOCK to be substantially more powerful than Structure in the absence of pure genotypes and that it had greater potential for fast processing. FLOCK is currently programmed in Maple. Programming of FLOCK in a compiled programming language such as Visual Basic or C++ will soon be undertaken to provide a friendlier interface and much shorter run times.

Acknowledgements

This work was motivated by an analysis of a beluga whale dataset for Mike Hammill (Department of Fisheries and Ocean of Canada). We are grateful to him for giving us the spare time to develop the FLOCK method. The work was also funded by the NSERC grant to J.T. We are also truly indebted to Craig Primmer who kindly provided his published microsatellite datasets on European grayling.

References

- Allinse C, Mariac C, Vigouroux Y *et al.* (2008) Role of seed flow on the pattern and dynamics of pearl millet (*Pennisetum glaucum* L. R. Br.) genetic diversity assessed by AFLP markers: a study in south-western Niger. *Genetica*, **133**, 167–178.
- Anderson E, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.
- Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.
- Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) *Genetix 4.05, logiciel sous Windows TM pour la génétique des populations*. Université de Montpellier II, Montpellier, France.
- Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.
- Bernatchez L, Wilson CC (1998) Comparative phylogeography of nearctic and palearctic fishes. *Molecular Ecology*, **7**, 431–452.
- Boyer MC, Muhlfeld CC, Allendorf FW (2008) Rainbow trout (*Oncorhynchus mykiss*) invasion and the spread of hybridization with native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*). *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 658–669.
- Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.
- Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.
- Duchesne P, Godbout M-H, Bernatchez L (2002) PAPA 2.0 (Package for the Analysis of Parental Allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes*, **2**, 191–194.
- Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution*, **18**, 672–675.
- Hansen MM (2002) Estimating the long-term effects of stocking domesticated trout into wild brown trout (*Salmo trutta*) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. *Molecular Ecology*, **11**, 1003–1015.
- Hansen MM, Bekkevold D, Jensen LF, Mensberg KLD, Nielsen EE (2006) Genetic restoration of a stocked brown trout *Salmo trutta* population using microsatellite DNA analysis of historical and contemporary samples. *Journal of Applied Ecology*, **43**, 669–679.
- Hedrick PW, Fredrickson RJ (2008) Captive breeding and the reintroduction of Mexican and red wolves. *Molecular Ecology*, **17**, 344–350.
- Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.
- Jacobsen F, Nesje M, Bachmann L, Lifjeld JT (2008) Significant genetic admixture after reintroduction of peregrine falcon (*Falco peregrinus*) in Southern Scandinavia. *Conservation Genetics*, **9**, 581–591.
- Koskinen MT, Sundell P, Piironen J, Primmer CR (2002) Genetic assessment of spatiotemporal evolutionary relationships and stocking effects in grayling (*Thymallus thymallus*, Salmonidae). *Ecology Letters*, **5**, 193–205.
- Maplesoft (1981–2004) *Maple*. Maplesoft, a division of Waterloo Maple Inc, Waterloo, Ontario, Canada.
- Mavarez J, Salazar CA, Bermingham E *et al.* (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, **441**, 868–871.
- Paetkau D, Calvert W, Sterling I, Strobeck C (1995) Microsatellite analysis of population-structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.
- Pella J, Masuda M (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin*, **99**, 151–167.
- Pella J, Masuda M (2006) The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 576–596.
- Perry WL, Feder JL, Dwyer G, Lodge DM (2001) Hybrid zone dynamics and species replacement between *Orconectes* crayfishes in a northern Wisconsin lake. *Evolution*, **55**, 1153–1166.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Randi E (2008) Detecting hybridization between wild species and their domesticated relatives. *Molecular Ecology*, **17**, 285–293.
- Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.
- Ryman N, Jorde PE (2001) Statistical power when testing for genetic differentiation. *Molecular Ecology*, **10**, 2361–2373.
- Seehausen O, Takimoto G, Roy D, Jokela J (2008) Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Molecular Ecology*, **17**, 30–44.
- 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.
- Sokal RR, Rohlf FJ (1981) *Biometry*, 2nd edn. W.H. Freeman, San Francisco, California.
- Taylor EB, Tamkee P, Sterling G, Hughson W (2007) Microsatellite DNA analysis of rainbow trout (*Oncorhynchus mykiss*) from western Alberta, Canada: native status and evolutionary distinctiveness of 'Athabasca' rainbow trout. *Conservation Genetics*, **8**, 1–15.
- Turgeon J, Bernatchez L (2001a) Mitochondrial DNA phylogeography of Lake ciscoe (*Coregonus artedii*): evidence supporting extensive secondary contacts between two glacial races. *Molecular Ecology*, **10**, 987–1001.

- Turgeon J, Bernatchez L (2001b) Clinal variation at microsatellite loci reveals historical secondary intergradation between glacial races of *Coregonus artedii* (Teleostei: Coregoninae). *Evolution*, **55**, 2274–2286.
- Turgeon J, Bernatchez L (2003) Reticulate evolution and phenotypic diversity in North American ciscoes, *Coregonus* ssp. (Teleostei: Salmonidae): implications for the conservation of an evolutionary legacy. *Conservation Genetics*, **4**, 67–81.
- Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Supporting information

Additional supporting information may be found in the online version of this article:

Fig. S1 (a) Run time and (b) number of iterations to reach stability as a function of N , the size of a randomly mixed sample of Lake ciscoes.

Fig. S2 Given the admixture building device (see text), the probability that a specimen be hybridized as a function of time as expressed by generation index.

Fig. S3 P_{anc} measures the similarity between the line dividing the original A and B groups at generation 0 and the line dividing the reference groups at later generations. Here it is represented as a function of generation index and the reference groups, based on six loci, are those obtained from FLOCK.

Table S1 Allele frequency distributions of artificial genotypes forming groups A and B used in simulations

Table S2 Spreadsheet showing an implementation of the calculation of the expected proportion (probability) of hybridized genotypes as a function of generation index. The results of those calculations are represented in Fig. S2.

Appendix S1 Detailed explanation of the calculation of the expected proportion of hybridized genotypes as a function of generation index

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix

Re-allocation number matrices

The sequence of re-allocation numbers obtained from iterating re-allocations is best represented in matrix form. For instance, given initial subsamples X^1 and Y^1 , the first column of the result matrix from the first re-allocation would correspond to the number of X^1 specimens being re-allocated to X^1 (first row) and to the number of X^1 specimens being re-allocated to Y^1 (second row). Similarly, the first and second rows of the second column would refer to numbers of Y^1 specimens re-allocated to X^1 and Y^1 , respectively.

The next X^i subsample, X^2 , would comprise all specimens re-allocated to X^1 , whether originally from X^1 or Y^1 . In other words, X^2/Y^2 corresponds to all specimens counted in the first/second row of the first re-allocation matrix. Re-allocation number matrices can of course be of any dimension.

The re-allocation process is exemplified with the Lake ciscoe example in Figure 1.

Number of re-allocations

How many times should the re-allocation process be applied? Clearly, there is no fixed answer to this question. Let us define state $E^i = X^i, Y^i, Z^i \dots$ = composition of the k subsamples at time i . In some cases, the sequence of E^i will

end up being a single state (a 'fixed point') looping as in $E^i E^i \dots$. In other cases, instead of a single state looping, several distinct states will repeat themselves as, for example, in $E^i E^j E^k E^i E^j E^k E^i E^j E^k \dots$. The $E^i E^j E^k$ string would then be referred to as an 'orbit'. Fixed points or orbits can be detected by examining the re-allocation numbers without actually identifying the members of the k subsamples.

Whenever a fixed point E^i is reached then obviously the last state to be considered is E^i .

If the procedure starts orbiting, then any state pertaining to the orbit may be considered since differences between orbit states will be insignificant.

Still in other cases, neither a fixed point nor an orbit can be detected but the *re-allocation number matrix* will be varying only slightly over several past re-allocations. Then the last state obtained should be the one retained.

We thereafter refer to the condition of having attained a fixed point, an orbit or a (nearly) stable allocation number matrix as 'stability'.

Once stability has been reached and some state E^F considered final, the corresponding k subsamples $X^F, Y^F, Z^F \dots$, the diagonal specimens of the last re-allocation matrix, should be taken as reference groups. In other words, the reference groups comprise those specimens that have remained in the same subsample during the last re-allocation (the one leading to E^F).