Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013

doi: 10.1111/2041-210X.12076

Inferring dispersal and migrations from incomplete geochemical baselines: analysis of population structure using Bayesian infinite mixture models

Philipp Neubauer[1]*†, Jeffrey S. Shima[1] and Stephen E. Swearer[2]

[1]*Victoria University Coastal Ecology Laboratory, School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand; and* [2]*Department of Zoology, University of Melbourne, Melbourne, Vic., Australia*

Summary

1. Geochemical and stable isotope tags are often used to attribute individual animals in a sample of mixed origins to distinct sources, be it spawning, overwintering or foraging habitats. In order for individuals to be uniquely classified to one source, modelling approaches generally assume that all potential sources have been characterized in terms of their geochemical signature. This assumption is rarely met in applications of geochemistry in environments where species distributions and spawning grounds are poorly known; statistical methods that can accommodate this problem are therefore essential.

2. We develop nonparametric Bayesian mixture models for geochemical signatures that estimate the most likely number of sources represented in a mixed sample, both in the absence and presence of baseline data. We then use a marginal clustering framework to evaluate the probability that a fish comes from a particular source.

3. Using both simulations and a previously analysed data set, we illustrate the method and highlight the potential merits and difficulties. These examples reveal how our interpretations of geochemistry data sets can change when potentially un-sampled sources are taken into account.

Key-words: Bayesian mixture, Dirichlet process, dispersal, geochemical tags, stock mixture

Introduction

Geochemical tags are routinely used to reconstruct migrations and estimate demographic connectivity of populations in both terrestrial and marine systems (Rubenstein & Hobson 2004; Elsdon *et al*. 2008). Trace element tags as well as stable isotopes contained within inert structures, such as fish otoliths, mollusc statoliths or bird feathers, have been used to quantify connectivity on ecological time-scales and study migratory pathways of a number of organisms including fish (Thorrold *et al*. 2001), birds (Rubenstein *et al*. 2002) and mammals (Burton & Koch 1999).

Inferences about geographical origins or migratory pathways from geochemical tags generally involve initial sampling of individuals from potential sources of interest (i.e. spawning, foraging or overwintering grounds) to establish a geographical baseline or reference atlas. Individuals of unknown origin are then assigned to one of the sources in this reference atlas based on their geochemical signature. The identifiability of potential sources is, intuitively, a major determinant of the success of such studies (Rubenstein & Hobson 2004; Elsdon *et al*. 2008). Furthermore, this approach generally requires, or at least (implicitly) assumes, that all potential sources have been sampled in order to determine geographical origins of individuals in the mixed sample. Omission of potential source sites can limit the inferences one can make regarding dispersal in a given system, since the assignment of individuals to a finite set of sources may be erroneous if one does not have a complete atlas. Because comprehensive sampling is often not feasible, the utility of this approach may be limited in many marine environments (Campana *et al*. 2000), although this may be of lesser concern if there is strong spatial covariation in signatures such that misassignments are made to spatially neighbouring sources (Pella & Masuda 2006; Munch & Clarke 2008).

Bayesian tools have gained considerable ground in the analysis of samples of mixed origin (i.e. stock mixtures in fisheries or individuals of unknown origin in connectivity and migration studies), in part, because they enable practitioners to define realistic and probabilistically sound models that can incorporate uncertainty at various levels of an analysis (Pella & Masuda 2006; Munch & Clarke 2008; Smith & Campana 2010; Pflugeisen & Calder 2012). In particular, Bayesian methods have been employed in an attempt to provide an answer to the problem of an unknown number of sources in a mixture. Pella & Masuda (2001) proposed posterior predictive checks for an unconditional (sensu Koljonen, Pella & Masuda 2005) Bayesian mixture model from genetic characteristics to identify potential mismatches between the baseline and the mixed source sample, which may indicate the presence of unsampled sources in the mixed sample (see also Smith & Campana 2010).

*Correspondence author. E-mail: neubauer.phil@gmail.com

†Present address: Dragonfly Science, PO Box 27535, Wellington 6141, New Zealand

It does not, however, provide a way to estimate the nature and number of such extra-baseline populations. White *et al.* (2008) proposed Bayesian model selection to find the most likely number of sources in a mixed stock, but this model does not explicitly connect the baseline with the mixed sample.

A Bayesian method to directly identify the contribution of extra-baseline sources to a mixture is provided by Pella & Masuda (2006). The use of a prior which has support over a theoretically infinite number of possible sources yields a marginal distribution over the number of sources in the mixed sample, thus eliminating the problem of model selection. Here, our aim is to develop and extend analogous models that directly infer the presence and contribution of extra-baseline sources, while following the distributional assumptions commonly employed for geochemical data. Statistical models are implemented in a new (open-source) package for Bayesian analysis of population structure implemented R (R Development Core Team 2007) and using the Julia language for technical computing to efficiently implement Markov Chain Monte Carlo methods for parameter estimation. The proposed models are tested on simulations as well as a well-known data set (weakfish: Thorrold *et al.* 2001), which has previously been evaluated by other authors to illustrate Bayesian methods for geochemistry data (Munch & Clarke 2008; White *et al.* 2008).

Statistical models

A DIRICHLET PROCESS MIXTURE MODEL FOR CLUSTERING

When the number of sources in a mixed sample is itself uncertain, identifying the (natal) origin of individuals becomes a difficult statistical problem: any classification will be biased by *a priori* exclusion of potential sources. The goal of the Dirichlet process mixture (DPM) introduced in this section is similar to the approach of White *et al.* (2008) in that we aim to infer the most parsimonious number of source populations in the sample. It represents a straightforward Bayesian extension of the finite mixture model that is commonly employed for classification in mixed sample analyses, whereby an analytical integration step makes it possible to circumvent the problem of model selection in mixture models (Celeux *et al.* 2006) and to infer the number of potential sources directly.

Practitioners working with geochemical tags commonly measure a suite of elements that are thought to be useful in discriminating potential sources. We assume that the p-dimensional vector of geochemistry data, denoted $y_i = y_{i,1},\ldots, y_{i,p}$, for fish i in a mixed sample $\boldsymbol{y} = y_i,\ldots, y_n$ of $i = 1, \ldots, n$ fish (all subsequent indexing proceeds in the same manner), is drawn from a multivariate normal distribution that characterizes the source of this individual (note that other distributions could be used here). Furthermore, a total of K separate sources are potentially represented in the mixed sample $\boldsymbol{y}$. In the case of a complete baseline reference atlas (hereafter, 'baseline'), a natural way to model these data is the finite mixture model, which models the joint density of the data $\boldsymbol{y}$ obtained from all n fish, each having originated from one of K sources.

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i|\theta_k) \qquad \text{eqn 1}$$

The set $\theta_k = \{\mu_k, \Sigma_k\}$ includes the mean vector and covariance matrix of the multivariate normal distribution, and $f(y_i|\theta_k)$ is thus the (p-dimensional) multivariate normal likelihood for y_i given the source parameters θ_k. A fish in the mixed sample thus originated from source k with probability $\pi_k = P(s_i = k)$, where s_i is a categorical variable assigning fish i to source k. Often, it is these probabilities that are of interest, for instance, when the focus is on stock mixing proportions in a fisheries context.

In a Bayesian context, one can conveniently write this model in a hierarchical fashion which directly illustrates conditional dependencies:

$$\begin{aligned} &y_i|\boldsymbol{\theta}, s_i \sim MVN(\theta_{s_i}) \\ &s_i|\boldsymbol{\pi} \sim MN(\boldsymbol{\pi}) \\ &\boldsymbol{\pi} \sim Dirichlet\left(\frac{\gamma}{K}\right) \\ &\theta_{s_i} \sim G_0 \end{aligned} \qquad \text{eqn 2}$$

Here, | denotes a conditional statement (a|b reads 'a given b') and ~ reads as 'is distributed as'. *MN* is the multinomial distribution and G_0 is the prior for the parameters of the multivariate normal (*MVN*) density in the first line, for instance a conjugate normal-inverse-Wishart prior (see, for instance, Gelman *et al.* 2003). For details about this prior and its parameters, see the Appendix S1 (Supporting Information). The Dirichlet distribution is the conjugate (natural) prior for the multinomial distribution, with concentration parameter γ.

When a complete baseline is unavailable or the assumption thereof is questionable, K will be unknown, and from a Bayesian perspective, K should then be treated as an uncertain parameter to be estimated. From the above hierarchical formulation, it is evident that only π depends explicitly on the choice of K, all other parameters, such as $\mathbf{s}$, only indirectly depend on K via π. Neal (1992) showed that, given a symmetrical Dirichlet prior for source probabilities π, one can integrate (1) with respect to π and take the limit of $K \to \infty$ to obtain a conditional prior for $\boldsymbol{s}$ that does not depend on K. The new prior, replacing lines two and three in the hierarchical formulation in eqn (2), can be written as follows:

$$\begin{aligned} &P(s_i = k|\boldsymbol{s}_{-i}) = \frac{n_k^{-1}}{n - 1 + \gamma} \\ &P(s_i \neq k \text{ for all } k \epsilon K|\boldsymbol{s}_{-i}) \end{aligned}$$

where $\boldsymbol{s}_{-i}$ is the vector of all source assignments, excluding individual i; n_k^{-i} is the number of individuals attributed to source k excluding individual i, or formally $\Sigma_{j \neq i}\delta(s_j = k)$, with $\delta(x)$ a point mass at x. Analogously to estimation of source probabilities $\boldsymbol{\pi}$ in the finite mixture model (c.f. Munch & Clarke 2008), this prior states that with a probability proportional to the number of individuals attributed to a given source, any individual of unknown origin will also have originated from that source. However, with a probability proportional to γ, this individual will have come from a previously uncharacterized

source. This model is often referred to as the infinite mixture model or a Dirichlet process mixture (DPM). To obtain Bayesian posterior source assignments, the prior that is formulated above is combined with the likelihood of belonging to a previously characterized or new source to give a posterior probability (after normalization in Bayes theorem). Specifically, the likelihood of belonging to a characterized source is MVN [as in (1)], with $\boldsymbol{\theta}$ estimated from all n^{-i} fish.

The prior for the concentration parameter γ determines how many sources are *a priori* expected to be sampled in the data set. A flat (uninformative) prior for γ would thus suggest that any number of sources from 1 to n are equally likely. Other priors may, however, be useful if the number of sources can be reduced to a range, or a 'best guess'. In this instance, some appropriate probability distribution that is defined on the positive real line (e.g. Poisson), or a specified vector of prior probabilities may be used. Since, however, a gamma distribution is the most natural prior for the concentration parameter γ, Dorazio (2009) suggested to use the Kullback-Leibler distance to match, as closely as possible, the parameters of the gamma distribution such that the resulting prior over the number of sources reflects the desired prior distribution. We implemented this approach to specify priors for the number of sources in the data set.

Further details about our implementation and Markov Chain Monte Carlo sampling algorithms can be found in the Appendix S1. A toolbox implemented in R implements (1) previously described (finite) Bayesian mixture analyses, (2) the new methods, and includes (3) pre- (e.g. priors) and postprocessing steps that are outlined below. This toolbox is available at http://github.com/Philipp-Neubauer/PopR, and readers are encouraged to contribute to this open-source software, to facilitate the eventual development of a comprehensive R toolbox for Bayesian analysis of population structure.

Visualizing model results

To visualize patterns in the mixed sample, we applied the exact linkage algorithm of Dawson & Belkhir (2009), which constructs a tree (akin to a hierarchical clustering tree) based on estimates of marginal co-assignment probability. This quantity expresses the probability that two individuals (or a set thereof) are assigned to the same source, given the uncertainty about the number and nature of these sources. It is a marginal probability: it integrates over the posterior distribution of model parameters and thereby takes into account the uncertainty inherent in their posterior distributions. This probability is also invariant to the label switching that accompanies iterative mixture model estimation and makes traditional (e.g. confusion matrix) representations difficult (Dawson & Belkhir 2009).

The node height of any node in the constructed tree is equal to the estimated posterior co-assignment probability of the individuals or sets of individuals which are merged by that node. It is worthwhile to note that the number of clusters separated by very low co-assignment probabilities (say $P_c < 0.05$) in the co-assignment tree is only expected to equal the most likely number of sources inferred by the model when there is no uncertainty in the marginal posterior distribution over the number of sources. To clarify this, one can imagine two of individuals from the mixed sample that are attributed to separate clusters when two sources are drawn in the model (during an MCMC iteration), but these sets are merged when only a single source is drawn (Fig. 1).

If it is twice as likely that the data originated from two sources as opposed to a single one, such that $(K = 2) = 2/3$ and $P(K = 1) = 1/3$, then the co-assignment probability P_c of these sets is 1/3. Thus, even though there are most likely two sources, the separation in terms of co-assignment probabilities is only 2/3. The co-assignment thus integrates over uncertainty in the number of sources to give a measure of separation of groups of fish (see Fig. 2 for an example of a tree constructed from a simulated data set).

If a baseline is present, this probability measures the marginal assignment probability of a fish or set of fish to a baseline source. The average co-assignment probability of a set of fish with a baseline or another set of individuals gives an estimate of the expectation of P_c.

Applications

We first conducted simulations to examine the properties of the DPM models in different idealized situations, as well as to test its robustness (see also the Appendix S1). We then tested the model on a data set of weakfish (*Cynoscion regalis*) otolith geochemistry, which was originally investigated by Thorrold *et al.* (1998, 2001). The premise of the original study was to investigate natal homing to five estuaries along the east coast of the United States (from north to south): New York (NY),

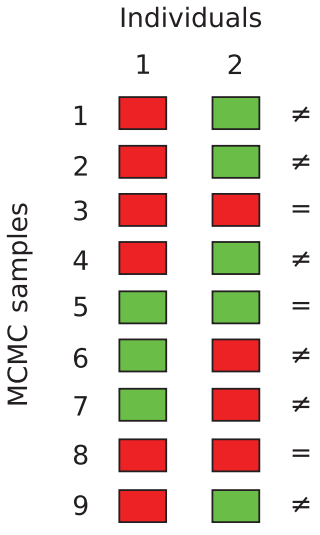

Fig. 1. Illustration of co-assignment probabilities from Markov Chain Monte Carlo estimation of mixture models. At each draw, individuals are attributed to a source (green or red), according to posterior assignment probabilities. For each sample (1–9), their source is either identical (=) or they are attributed to distinct sources (≠). In this case, the probability that the two individuals belong to the same source is $P_c = 3/9 = 1/3$, and the probability that the two individuals come from distinct sets is $1\text{-}P_c = 6/9 = 2/3$.

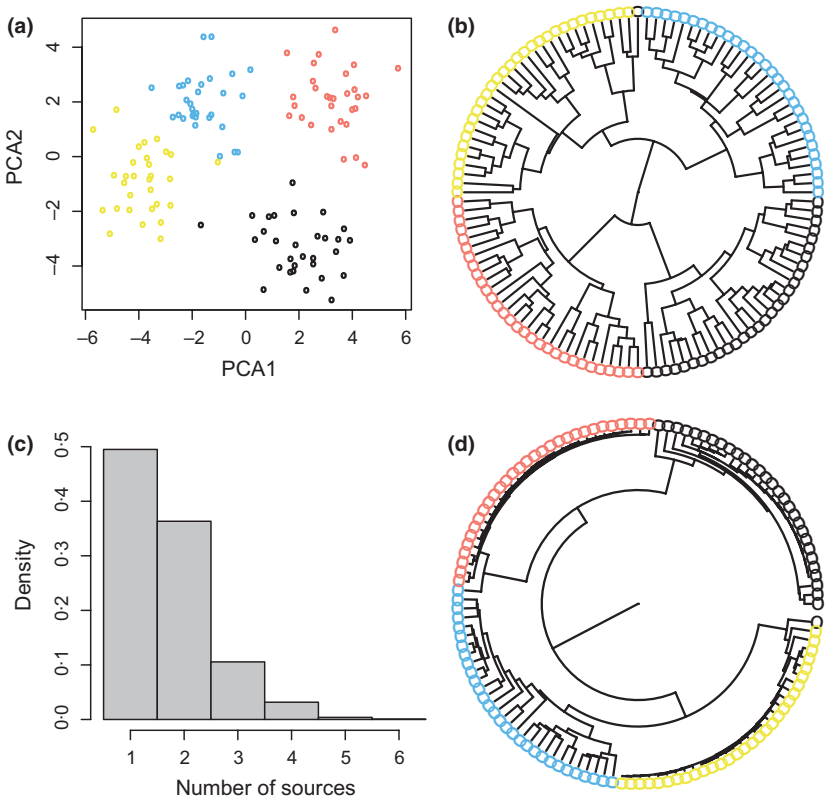

Fig. 2. Simulated example illustrating an application of the DPM without a baseline for a data set drawn from four well-separated sources. Panels a-d illustrate (a) the data (colour-coded throughout to represent individual source) projected on the first two principal components, (b) a hierarchical clustering of the data, (c) the posterior distribution of the number of sources and (d) tree of posterior co-assignment probabilities. In exact linkage tree (d), the centre of the circular tree suggests a co-assignment probability of zero, with co-assignment increasing concentrically towards the periphery, where the co-assignment probability is 1. Individuals (circles on tree leaves) mainly group into the correct sources, separated by low co-assignment probabilities. Only for the source that is intermediate in terms of its signatures (yellow source) is there greater uncertainty reflected by low co-assignments within the source relative to the other sources.

Delaware Bay (DE), Chesapeake Bay (CB), North Carolina (NC) and Georgia (GA). The authors used otolith core geochemistry of adult weakfish which were compared with a baseline of geochemical signatures collected from juvenile weakfish 2 years earlier. A discriminant analysis was used to assign adults to natal estuaries, thus assuming that (i) adult fish were spawned in one of five baseline estuaries and (ii) that these estuaries were sufficiently characterized by the data collected from juvenile fish to allow for such a classification.

There are a number of motivations for using DP models for a re-analysis of this data set. (i) Recent models developed for assigning natal origins (or estimating source proportions) based on otolith geochemistry have used this data set for illustrative purposes (Munch & Clarke 2008; White *et al.* 2008), (ii) While the original study notes that the five estuaries under investigation account for 90% of commercial weakfish catch, other estuaries along the east coast could have potentially accounted for some of the adult fish in this study, (iii) Some of these estuaries are relatively large (i.e. CB is the largest estuary in the USA), and large variation in otolith geochemistry can often be found even on small scales within estuaries (Thorrold *et al.* 1998; Gillanders & Kingsford 2000; Miller 2007). Since fish in the mixed sample may originate from unsampled locations within these estuaries, this could lead to uncertainty in inferences about the strength of natal homing as derived from geochemical tags.

Clustering fish without a baseline

In the absence of a baseline, the DPM model could, in theory, be used to estimate the number of distinct sources that make up the mixed samples. In that case, the likelihood for y_i is simply the likelihood used in (1), with $\boldsymbol{\theta}$ estimated from the mixed sample itself. Since we estimate source parameters from data, posterior source distributions have heavier tails than a normal distribution [namely, the posterior source distributions are generalized (multivariate) Student's t-distributions, see also (Munch & Clarke 2008)], and the procedure should thus be robust to deviation from the normal assumption in that sense.

In practice, estimating the number of sources in a multivariate data set is a difficult task at best and is only a sensible thing to do if sources are rather well separated in parameter space – in which case source partitioning should be rather obvious from, say, a plot of the principle components of the data (see Fig. 2 for a simple, illustrative simulation example). More realistic simulations show that it is difficult to obtain interpretable

results without a baseline. In the Appendix S1, we provide more details about simulations and discussion about the difficulty in estimating the number of sources in realistic settings and aim to briefly describe the nature of the problem here.

The difficulty of identifying the number of sources increases with the dimensionality of the data: whereas a higher number of (informative) geochemical signatures (chemical elements) facilitates discrimination among sources (e.g. Neubauer *et al.* 2010), it also leads to an exponentially sparser distribution of the samples in multivariate space [the well-known *Curse of Dimensionality*, e.g. Bishop (2007)]. Thus, in order to maintain a sample that adequately characterizes the (multivariate normal) distribution of sources in the data set, the number of individual samples needs to increase exponentially with an increasing number of signatures. Such a concomitant increase in sample size is generally not feasible in practice. If a low number of principal axes account for most of the variability in the data, statistical approaches (e.g. principal component analysis) may be implemented to reduce dimensionality and increased performance of the DPM for the identification of individual sources. In other instances, informative priors [e.g. describing source (co)variances] may be used to help characterize the nature of the source-specific distributions. In practice, however, this approach would normally require an informative baseline.

Using the DPM with a baseline of normally distributed sources

This application uses the DPM in a set-up similar to that of a classical finite mixture model, in which geochemical signatures from each source are assumed to be normally distributed. In contrast to the classical mixture model, however, we do not need to assume that only the sampled sources contribute to the mixed sample. The vector of source parameters $\boldsymbol{\theta}$ is now estimated from the baseline alone (conditional, e.g. Munch & Clarke 2008) or jointly from the baseline and the mixed sample (unconditional, e.g. Smith & Campana 2010). The latter case is generally preferred (Koljonen, Pella & Masuda 2005) and is implemented here. The new likelihood for y_i being attributed to source k is then conditional on the baseline samples $\boldsymbol{x}$ (and $\boldsymbol{y_k^{-i}}$ for unconditional classification) and can be written as follows:

$$f\left(y_i|\theta_k, \boldsymbol{x}, \left(\boldsymbol{y_k^{-i}}\right)\right) = p(y_i|\theta_k)p\left(\theta_k|\boldsymbol{x}, \left(\boldsymbol{y_k^{-i}}\right)\right).$$

Again, an explicit integration over θ yields

$$f\left(y_i|\boldsymbol{s}\theta_i = \boldsymbol{k}, \boldsymbol{x}\left(\boldsymbol{y_k^{-i}}\right)\right) = \int p(y_i|\theta_k)p\left(\theta_k|\boldsymbol{x}, \left(\boldsymbol{y_k^{-i}}\right)\right)d\theta_k,$$

an expression for the likelihood that integrates over uncertainty in $\boldsymbol{\theta}$, resulting in a density with larger tails. Further details about the nature and parameters of this predictive density can be found in the Appendix S1.

In the presence of a baseline, we have information about the parameters of the normal distributions that characterize the sampled baseline. Thus, as long as the distributional assumptions hold approximately and sources are reasonably well resolved by the geochemistry, one should expect the model to assign fish to a characterized or an uncharacterized baseline source according to the posterior probability of each. The second difference with respect to the first application is that, in this case, we can more reliably estimate the parameters of G_0 (the prior for source characteristics) since we now have information about the characteristics of at least some sources and can formulate our prior accordingly. Since the probability that an individual comes from a previously uncharacterized source is directly linked to G_0, this will be crucial to making reasonable inference about the presence of extra-baseline sources.

A simulated example (Fig. 3a) illustrates how the model classifies individuals to baseline (i.e. sources identified in a reference atlas) and extra-baseline sources (i.e. sources that are not represented in the reference atlas), in a relatively realistic setting (simulated sources overlap in parameter space, but remain relatively well discriminated – Fig. 3a). Since we have sampled a baseline of four sources, we hypothesize that there are around four sources in our mixed sample: after all, the sampling effort for the baseline should mimic our hypothesis about dispersal/migrations. We thus emulate (via the prior on γ, as discussed above) a Poisson distributed prior for the number of

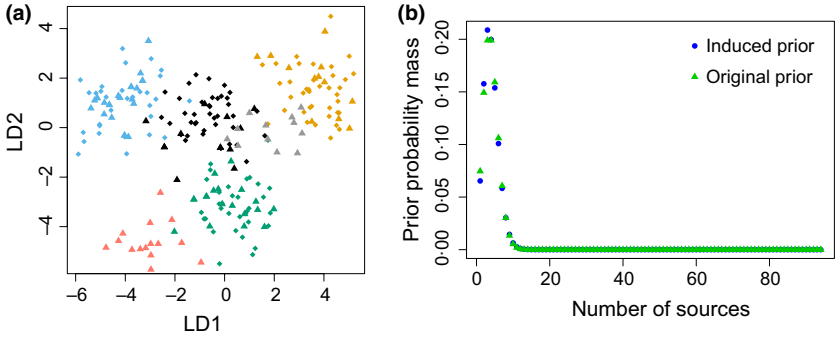

Fig. 3. (a) Simulated baseline, with colours indicating source membership, diamonds indicating the four sampled baseline sources and triangles depicting the mixed sample, including individuals from two extra-baseline sources (grey and red) that appear only in the mixed sample. (b) Original Poisson (4) prior (green triangles) and the induced prior on the number of sources (blue circles) by using a Gamma (*a*,*b*) distribution prior on the DPM concentration parameter γ to reflect the original (explicit) prior, with parameters *a* and *b* optimized to match the original and induced distributions as closely as possible.

sources with mean 4 (and hence variance 4 – Fig. 3b) that puts most of the prior weight on observing 4 or fewer sources [$P(K \leq 4) = 0.63$], but allows, with probability $P(K > 4) = 0.37$, for more sources to be observed.

The priors for source parameters given by G_0 are set to the harmonic mean of the source covariance matrices for the prior covariance matrix, and we set the prior degrees of freedom (reflecting certainty about this value) to the (algebraically) allowed minimum of $p + 1$, thus expressing relatively low confidence in this value. The prior mean and a scaling parameter for the covariance (k_0 – see the Appendix S1 for details) are estimated from the baseline and mixed sample, using a flat prior for the former and a Gamma distribution with scale 1 and shape 1 as hyperparameters for the latter.

Model results suggest that, given a reasonably resolved baseline as in this example, the model can recover extra-baseline sources and provide reasonable assignment success even under only vaguely informative priors (Fig. 4). The two simulated extra-baseline sources have a co-assignment probability of 0 with the baseline sources and a probability of 0.07 of belonging to the same source. Some misclassification is nonetheless obvious: mixed samples from the black source are assigned to their baseline with an average P_c of only 0.15. Some of these mixed samples are attributed to an extra-baseline source, showing that the DPM method can have a drawback in that mixed samples coming from baseline sources may be erroneously attributed to extra-baseline if they fall within the tails of a distribution. This should be especially prevalent if sources are not adequately sampled, and their estimated variance for each element is lower than the actual variance.

We conducted another set of simulations to test our methods on the weakfish data set. For these, we used the baseline to conduct two relevant tests. Specifically, we asked (1) whether sources are omitted from the baseline and treated as a mixed sample, are these consistently assigned to an extra-baseline source and (2) do random samples from the baseline get assigned to their respective source, and not to extra-baseline samples. To test (1), we omitted each source in turn from the baseline and computed the co-assignment probabilities P_S with each remaining source for all samples, which allowed us to calculate the probability of coming from an extra-baseline source as $P_{EB} = 1 - P_S$. To investigate question (2), we used 20 cross-validation trials of 30 random samples from the baseline as a mixed sample, and the remainder of the baseline was kept as the new baseline. For each trial, we calculated the proportion of individuals correctly assigned to their respective sources, as well as the probability that any of them originate from an extra-baseline source. The prior emulated a negative binomial distribution with mean 4 (5) and rate 4 (5) as the prior for the number of sources in the mixed sample for test i (ii), leading to $P[K \leq 4(5)] = 0.64$ (0.62). All other priors were calculated as for the simulated example. We lastly compared these cross-validations to finite mixtures, using both conditional assignment (Munch & Clarke 2008) and unconditional assignments (Smith & Campana 2010), using the same prior as for the DPM model.

The tests confirm that the model performs well when assigning samples to extra-baseline sources in test (i), with very few samples having co-assignment probabilities $P_{EB} < 1$ (Fig 5). The DPM generally assigned individuals to their correct estuaries in test (ii), albeit with lower probabilities than the conditional and unconditional assignments. This was due to nonzero probabilities that individuals originated from extra-baseline sources (Fig. 5). This uncertainty is obviously eliminated in finite mixtures by the strong hypothesis that we know the number of sources. Probabilities for extra-baseline sources were, however, generally low enough (and well below co-assignment probabilities with baseline sources) such that very low proportions of samples were erroneously assigned to extra-baseline sources. Nevertheless, it is useful to note that the possibility of observing extra-baseline sources generally leads to lowered assignment probabilities overall unless the probability of additional sources is very low.

To analyse the weakfish data set, we initially used the same priors as above, this time emulating a negative binomial distribution with mean 5 and rate 5 as the prior for the number of sources in the mixed sample (Fig. 6a), leading to $P(K \leq 5) = 0.62$. This seems both flexible and reasonable since a

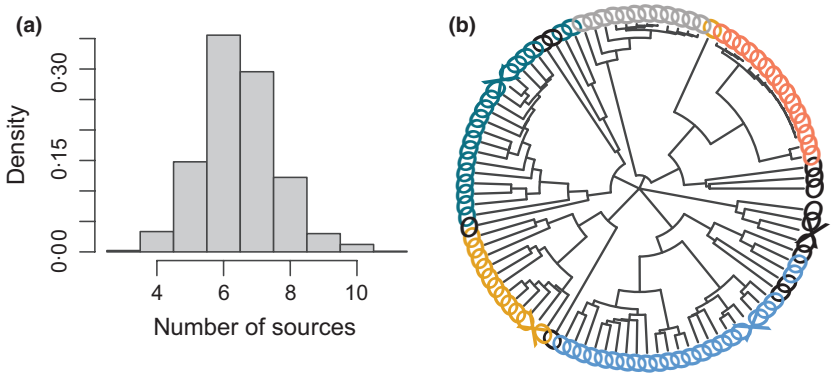

Fig. 4. Posterior distribution over the number of sources and posterior co-assignment trees for simulated data depicted in Fig. 3. Rounds on leaves are mixed sample individuals, which are clustered with the baseline sources (crosses on leaves) according to their co-assignment probabilities with the sources. Since co-assignment for the baseline samples within a source is always one (their source identity is known), baseline samples are collapsed to one leaf per source in the tree.

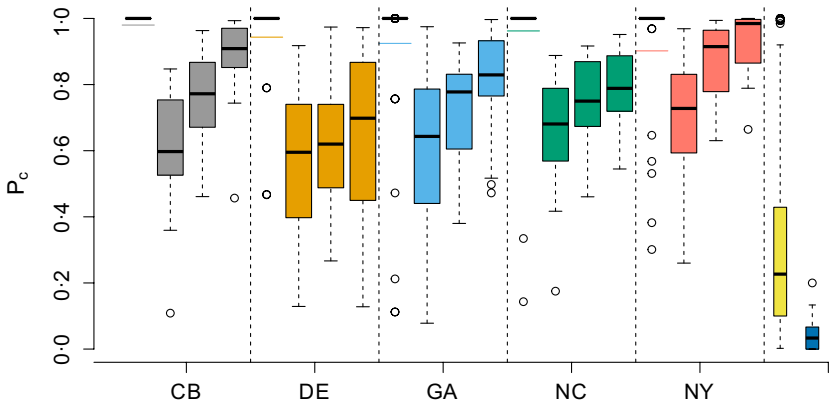

Fig. 5. Testing the DPM model with the weakfish baseline samples from different estuaries (CB = Chesapeake Bay, DE = Delaware Bay (DE); GA = Georgia; NC = North Carolina; NY = New York). For each estuary, four types of information are shown. (i) Leftmost boxplots for each focal estuary (in black, error bars obscured by most boxes) give the estimated probability of coming from an extra-baseline source, when the focal estuary was excluded from the analysis (the coloured lines beneath these boxplots give the proportion of individuals that were assigned to an extra-baseline source in this scenario). The second, third and fourth boxplots for each focal estuary give the assignment probabilities to the correct baselines for (ii) the DPM, (iii) conditional and (iv) unconditional finite mixtures. The last two columns of the figure show boxplots of individuals' assignment probabilities (yellow) and actual assigned proportion (blue) of samples to extra-baseline sources in Trial 2.

high degree of natal homing was previously found, but straying between estuaries was also inferred from the geochemistry. Nevertheless, the majority of Weakfish spawners are likely found within the sampled estuaries, warranting a high probability that $K \leq 5$.

Our analysis of the adult weakfish data set strongly suggests that the juvenile weakfish baseline does not provide a sufficient basis to assign weakfish adults to spawning estuaries in many cases (Fig 6): all but one of the North Carolina samples (Fig. 6e) and the majority of New York samples (Fig. 6f) group away from the baseline as extra-baseline sources. The expected co-assignment of adults collected in these estuaries with the corresponding juvenile signatures is thus 0 and 0.15, respectively, suggesting that we cannot reliably conclude that natal homing is prevalent among adults that are found in these estuaries. For the remaining three estuaries, the majority of adult fish group most closely with their respective baseline. Nevertheless, in both Chesapeake and Delaware Bay (Fig. 6b, c), the expected assignment probabilities to the baseline remain relatively low, with 0.55 and 0.43, respectively. Only in Georgia (Fig. 6d) do we see a clear association of the weakfish adults collected in that estuary with the corresponding baseline samples, with $P_c = 0.88$. Our analysis provides only very limited support for the straying or dispersal among estuaries that was suggested by the previous analyses of this data set. In fact, only in North Carolina and Georgia do we find individual fish (1 and 4 fish, respectively) that group more closely with other baselines than with their own or with extra-baseline samples.

Discussion

We have developed a modelling approach that addresses a long-standing challenge associated with the application of geochemical signatures to estimate dispersal and migration, namely the uncertainty in the number of contributing sources. Clustering and classification procedures based on our methods may thus provide considerable insight into patterns in geochemical data, both at the level of the baseline and the mixed sample. Since these methods can be directly derived from finite mixtures (Neal 1992), classification of individuals to specific sources will be the same (with same accuracy) as in a finite mixture model when the baseline is complete and representative of the mixed sample. However, the possibility for recruits to come from un-sampled sites makes the DPM approach an excellent tool for exploring and inferring connectivity based on geochemical tracers in geographically and chemically complex landscapes. The DPM model is thus a more realistic procedure in most applications of geochemical tags in the marine environment, where the requirement of a complete baseline can rarely be achieved.

Our re-analysis of the weakfish data set from Thorrold *et al.* (2001) provides an illustration of such challenges in a coastal environment. We find evidence for extra-baseline sources in all mixed samples except those collected in GA. In NC and NY, the majority of fish group in such extra-baseline clusters, and we can conclude very little about natural homing from these two estuaries. While for GA there is little uncertainty about natal homing, there is considerably greater uncertainty about this process for both the CB and DE estuaries despite the majority of mixed grouping with the baseline of their collection estuary. As for our trials with the weakfish analysis, this was due to the possibility for extra-baseline sources in both cases, and the co-probabilities with the collection estuary's baseline were not far below those found in cross-validation trials.

There are at least three (nonexclusive) possible explanations for the lack of association between baseline and mixed sample in both NC and NY, as well as DE and CB to a lesser degree. The first is that this pattern is indicative of extra-baseline sources that could not be found with traditional analysis methods. However, the authors of the original study point out that over 90% of weakfish are caught in one of these five estuaries, and it is thus unlikely

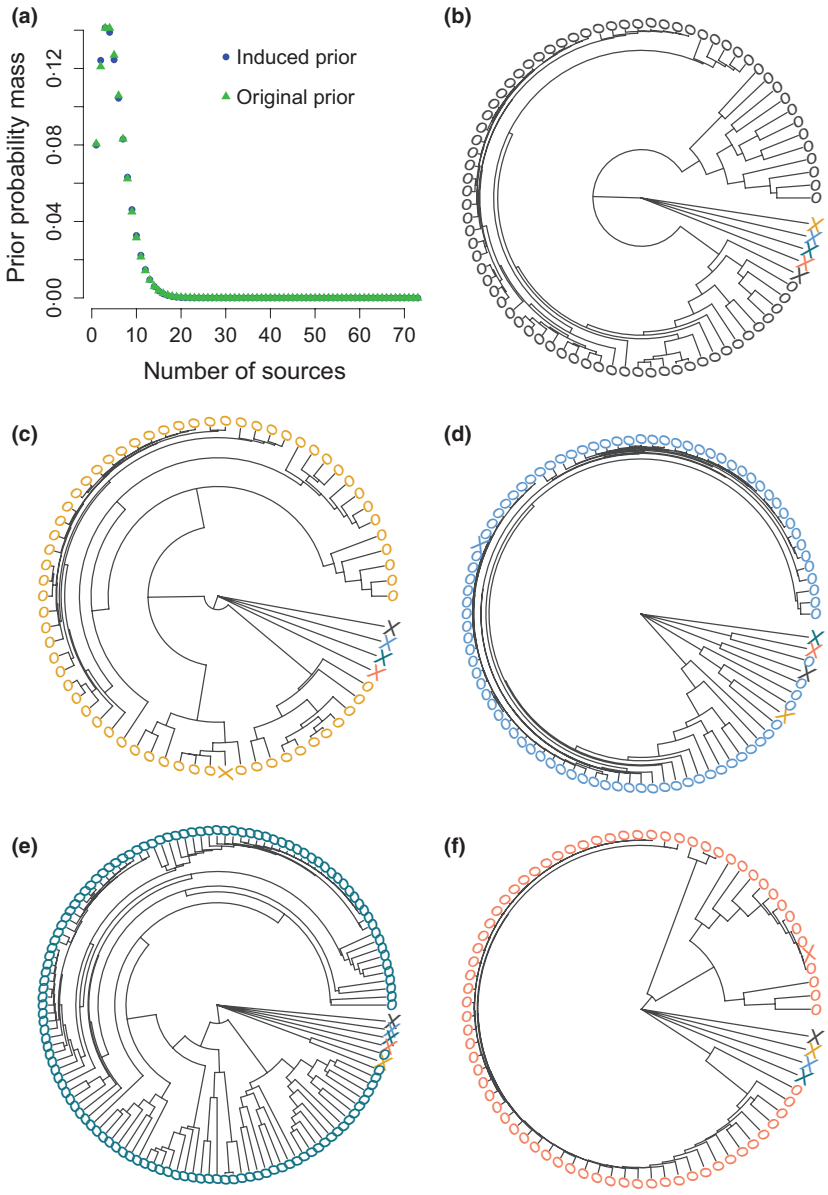

Fig. 6. (a) Example prior for weakfish analysis using a negative binomial distribution. (b–f) Adult weakfish exact linkage trees from the DPM with baseline (juvenile weakfish) for each collection estuary: (b) Chesapeake Bay, (c) Delaware Bay, (d) Georgia, (e) North Carolina and (f) New York. Rounds on leaves are mixed sample individuals (adult weakfish), and crosses on the leaves designate the (collapsed) baseline (see Fig. 4 for detail).

(though no impossible) that most individuals in major estuaries are spawned in other minor estuaries.

A second possibility is that geochemical signatures are spatio-temporally heterogeneous within estuaries, and the baseline sample from each estuary is not entirely representative of the variability of signatures therein, meaning that extra-baseline sources are likely to originate from unsampled locations within one of the sampled estuaries. Indeed, earlier studies reported significant differences between locations within estuaries (Thorrold *et al.* 1998), and significantly different source locations within each estuary may have contributed to the actual mixed sample, increasing overall source variances relative to the baseline. Temporal variability in signatures may have further contributed to the lack of association.

A third and entirely different potential reason for the low co-assignment of mixed and source samples is the different sampling methods used to collect chemical signatures for the juvenile baseline and the adult mixed sample (solution-based ICPMS for the juvenile baseline and laser ablation ICPMS for adult samples). Though the data were standardized to make baseline and mixed samples comparable, this procedure may introduce a bias that is important enough to drive the DPM model to consider extra-baseline sources as the most likely origin for the majority of mixed samples.

Regardless of the explanation for this pattern, mixed sample signatures for both NC and NY in our analysis were sufficiently different from any juvenile samples, and previous classifications using assumptions of known baselines may have resulted in a misclassification of these individuals by assigning

them to the 'nearest' source in parameter space. This example thus aptly illustrates how the assumption of a complete baseline can mask uncertainty in the inference of natal homing. For strong inference of natal homing, one could aim for more extensive sampling to ensure that the baseline variability is representative of the variability found over the geographical extent defined as a 'source'.

Our simulations and trials with real data show that the number of sources in a sample can often be estimated using the DPM when a baseline is present to estimate source-specific parameters (and when assumptions are approximately met). Yet, it is also evident that this task is increasingly difficult without a baseline or when the baseline sources are not entirely representative of source locations, either because they are undersampled (e.g. significant within source variation and few baseline samples) or due to methodological discrepancies. As with the weakfish data, it may often be difficult to know with certainty if inferred extra-baseline sources are real or due to insufficiently sampled variability at the baseline level.

The DPM performs a clustering of the data while building a distribution over the number of sources. Without a baseline, it is difficult to ensure that the geographical scale of a source corresponds to a unimodal (normal) distribution that is the basis for clusters formed in the DPM and that individual sources are well separated (see also the Appendix S1). The biggest limitation of the DPM and any clustering method without a baseline is thus the interpretability of results on a geographical scale (see White *et al*. 2008 for a detailed discussion). Nevertheless, the DPM modelling approach has several advantages over existing methods for discovering structure in a recruit pool based on geochemistry. Current methods that use geochemistry to uncover the number of sources in a recruit data set or a mixed fishery use model selection or resampling criteria to produce a single best model (White *et al*. 2008; Fontes *et al*. 2009; Shima & Swearer 2009). The DPM model produces a marginal distribution over the number of sources, the direct probabilistic interpretation of which is more natural than that of arbitrarily scaled model selection criteria such as the AIC or DIC, and allows for estimation of marginal quantities such as $P(K^+ > S|M)$, the probability that there are more sources in the mixed sample than in the baseline S, given the specified model M. Furthermore, no previous approaches for eliciting the number of sources explicitly incorporate the baseline into the analysis – for the Bayesian clustering model of White *et al*. (2008), the geographical origin of clusters needs to be determined by comparison of cluster means of mixed sample and baseline fish. The DPM approach on the other hand can incorporate the baseline directly, while allowing for additional extra-baseline sources. The marginal description of relatedness of individual fish and clusters, expressed by the co-assignment probabilities, integrates over this uncertainty in the number of contributing sources and allows for a more thorough exploration of the structure of the mixture and the baseline.

The methods developed in this study can be extended to include genetic characters by combining it with the model proposed by Pella & Masuda (2006) [i.e. the likelihood for a source becomes the product of genetic and geochemistry likelihoods (Smith & Campana 2010)]. They can also be extended to many-to-many analyses as detailed in Bolker *et al*. (2007), which simultaneously models multiple mixed samples in an unconditional analysis. Finally, the DPM model could be applied to other types of geochemical tracers, such as stable isotopes, and other forms of data that are assumed to follow a normal distribution.

Like any modelling approach, the insights that can be gained from this method are only as good as the data itself and the model's appropriateness for this data. We encourage practitioners to compare results from our proposed approach to those from conditional and unconditional versions of the finite mixture model, which are supplied with the DPM software, and to critically evaluate discrepancies between outcomes, which may point to future research needs. Thinking carefully about the assumptions in each framework will allow practitioners to make sensible inferences in the light of assumptions deemed most appropriate for a particular data set. With a reasonable prior and assumptions holding approximately, our approach is shown to perform well with regard to estimating missing sources in the presence of a baseline data set, which to date, has been a significant limitation for the interpretation of otolith geochemical signatures in complex environments. The DPM methods, in concert with previous Bayesian methods for geochemistry data provided with our software, provide a considerable step towards a more flexible and realistic approach for analysing geochemistry data in studies of dispersal and migration.

Acknowledgements

This work was funded by a Royal Society of New Zealand Marsden grant to J.S.S. and S.E.S. which supported a PhD fellowship for P.N. The first author thanks Michele Masuda, Russel Millar and Will White for comments on early drafts that improved the manuscript. The authors thank Simon Thorrold for kindly sharing his data, and the anonymous referees whose exceptional reviews considerably improved previous submissions.

References

Bishop, C. (2007) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, New York.

Bolker, B.M., Okuyama, T., Bjorndal, K.A. & Bolten, A.B. (2007) Incorporating multiple mixed stocks in mixed stock analysis: 'many-to-many' analyses. *Molecular Ecology*, **16**, 685–695.

Burton, R.K. & Koch, P.L. (1999) Isotopic tracking of foraging and long-distance migration in northeastern Pacific pinnipeds. *Oecologia*, **119**, 578–585.

Campana, S.E., Chouinard, G.A., Hanson, J.M., Frechet, A. & Brattey, J. (2000) Otolith elemental fingerprints as biological tracers of fish stocks. *Fisheries Research*, **46**, 343–357.

Celeux, G., Forbes, F., Robert, C.P. & Titterington, D.M. (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–673.

Dawson, K.J. & Belkhir, K. (2009) An agglomerative hierarchical approach to visualization in Bayesian clustering problems. *Heredity*, **103**, 32–45.

Dorazio, R.M. (2009) On selecting a prior for the precision parameter of Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, **139**, 3384–3390.

Elsdon, T., Wells, B., Campana, S., Gillanders, B., Jones, C., Limburg, K. *et al*. (2008) Otolith chemistry to describe movements and life-history parameters of fishes. *Oceanography and Marine Biology: An Annual Review*, **46**, 297–330.

Fontes, J., Caselle, J.E., Sheehy, M.S., Santos, R.S. & Warner, R.R. (2009) Natal signatures of juvenile Coris julis in the Azores: investigating connectivity scenarios in an oceanic archipelago. *Marine Ecology Progress Series*, **387**, 51–59.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2003) *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman & Hall, Boca Raton.

Gillanders, B.M. & Kingsford, M.J. (2000) Elemental fingerprints of otoliths of fish may distinguish estuarine 'nursery' habitats. *Marine Ecology Progress Series*, **201**, 273–286.

Koljonen, M.L., Pella, J.J. & Masuda, M. (2005) Classical individual assignments versus mixture modeling to estimate stock proportions in Atlantic salmon (Salmo salar) catches from DNA microsatellite data. *Canadian Journal of Fisheries and Aquatic Sciences*, **62**, 2143–2158.

Miller, J.A. (2007) Scales of variation in otolith elemental chemistry of juvenile staghorn sculpin (Leptocottus armatus) in three Pacific Northwest estuaries. *Marine Biology*, **151**, 483–494.

Munch, S.B. & Clarke, L.M. (2008) A Bayesian approach to identifying mixtures from otolith chemistry data. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 2742–2751.

Neal, R. (1992) Bayesian mixture modeling. *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, 1991* (eds C.R. Smith, G.J. Erickson & P.O. Neudorfer), pp. 197–211. Kluwer Academic Publishers, Dordrecht.

Neubauer, P., Shima, J.S. & Swearer, S.E. (2010) Scale-dependent variability in *Forsterygion lapillum* hatchling otolith chemistry: implications and solutions for studies of population connectivity. *Marine Ecology Progress Series*, **415**, 263–274.

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.

Pflugeisen, B. & Calder, C. (2012) Bayesian hierarchical mixture models for otolith microchemistry analysis. *Environmental and Ecological Statistics*, **20**, 179–190.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubenstein, D.R. & Hobson, K.A. (2004) From birds to butterflies: animal movement patterns and stable isotopes. *Trends in Ecology & Evolution*, **19**, 256–263.

Rubenstein, D.R., Chamberlain, C.P., Holmes, R.T., Ayres, M.P., Waldbauer, J.R., Graves, G.R. *et al*. (2002) Linking breeding and wintering ranges of a migratory songbird using stable isotopes. *Science*, **295**, 1062–1065.

Shima, J.S. & Swearer, S.E. (2009) Larval quality is shaped by matrix effects: implications for connectivity in a marine metapopulation. *Ecology*, **90**, 1255–1267.

Smith, S.J. & Campana, S.E. (2010) Integrated stock mixture analysis for continuous and categorical data, with application to genetic-otolith combinations. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 1533–1548.

Thorrold, S., Jones, C., Swart, P. & Targett, T. (1998) Accurate classification of juvenile weakfish *Cynoscion regalis* to estuarine nursery areas based on chemical signatures in otoliths. *Marine Ecology Progress Series*, **173**, 253–265.

Thorrold, S.R., Latkoczy, C., Swart, P.K. & Jones, C.M. (2001) Natal homing in a marine fish metapopulation. *Science*, **291**, 297–299.

White, J.W., Standish, J.D., Thorrold, S.R. & Warner, R.R. (2008) Makov Chain Monte Carlo Methods for assigning larvae to natal sites using natural geochemical tags. *Ecological Applications*, **18**, 1901–1913.

Received 10 January 2013; accepted 27 May 2013
Handling Editor: Luca Borger

Supporting Information

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

Appendix S1. Methods and simulation details.