# Uncertainty in empirical estimates of marine larval connectivity

David M. Kaplan, Marion Cuif, Cécile Fauvelot, Laurent Vigliola, Tri Nguyen-Huu, Josina Tiavouane, Christophe Lett

ICES Journal of Marine Science

ICES Journal of Marine Science (2017), 74(6), 1723–1734. doi:10.1093/icesjms/fsw182

**Contribution to the Themed Section: '*Beyond ocean connectivity: new frontiers in early life stages and adult connectivity to meet assessment and management*'**

## Original Article

# Uncertainty in empirical estimates of marine larval connectivity

David M. Kaplan,[1]* Marion Cuif,[2,3] Cécile Fauvelot,[3] Laurent Vigliola,[3] Tri Nguyen-Huu,[4] Josina Tiavouane,[3] and Christophe Lett[2‡]

[1]*Virginia Institute of Marine Science, College of William & Mary, PO Box 1346, 1375 Greate Road, Gloucester Point, VA 23062-1346, USA*
[2]*Sorbonne Universités, UPMC Univ Paris 06, IRD, Unité de modélisation mathématique et informatique des systèmes complexes (UMMISCO), Bondy F-93143, France*
[3]*Institut de Recherche pour le Développement (IRD), Laboratoire d'Excellence LABEX Corail, UMR ENTROPIE, Nouméa cedex BP A5 98848, New Caledonia*
[4]*UMI IRD 209 UPMC UMMISCO, IXXI, 7 rue du Vercors, Lyon 69007, France*

**Corresponding author: tel: +1-804-684-7005; e-mail: dmk@vims.edu*
[‡]Present address: UMR MARBEC, Station Ifremer de Sète, Avenue J. Monnet, Sète 34203, France Cedex.

Kaplan, D. M., Cuif, M., Fauvelot, C., Vigliola, L., Nguyen-Huu, T., Tiavouane, J. and Lett, C. Uncertainty in empirical estimates of marine larval connectivity. – ICES Journal of Marine Science, 74: 1723–1734.

Received 9 March 2016; revised 23 September 2016; accepted 26 September 2016; advance access publication 26 December 2016.

Despite major advances in our capacity to measure marine larval connectivity (i.e. the pattern of transport of marine larvae from spawning to settlement sites) and the importance of these measurements for ecological and management questions, uncertainty in experimental estimates of marine larval connectivity has been given little attention. We review potential uncertainty sources in empirical larval connectivity studies and develop Bayesian statistical methods for estimating these uncertainties based on standard techniques in the mark-recapture and genetics literature. These methods are implemented in an existing R package for working with connectivity data, ConnMatTools, and applied to a number of published connectivity estimates. We find that the small sample size of collected settlers at destination sites is a dominant source of uncertainty in connectivity estimates in many published results. For example, widths of 95% CIs for relative connectivity, the value of which is necessarily between 0 and 1, exceeded 0.5 for many published connectivity results, complicating using individual results to conclude that marine populations are relatively closed or open. This "small sample size" uncertainty is significant even for studies with near-exhaustive sampling of spawners and settlers. Though largely ignored in the literature, the magnitude of this uncertainty is straightforward to assess. Better accountability of this and other uncertainties is needed in the future so that marine larval connectivity studies can fulfill their promises of providing important ecological insights and informing management questions (e.g. related to marine protected area network design, and stock structure of exploited organisms). In addition to using the statistical methods developed here, future studies should consistently evaluate and report a small number of critical factors, such as the exhaustivity of spawner and settler sampling, and the mating structure of target species in genetic studies.

**Keywords:** connectivity, larval dispersal, parentage analysis, self-recruitment, transgenerational marking.

## Introduction

Larval dispersal plays a critical role in the population dynamics of many marine species (Botsford *et al.*, 2009). In particular, the study of larval connectivity is central to marine spatial planning (Sala *et al.*, 2002; Botsford *et al.*, 2003), the identification of demographically separated fish populations (Garavelli *et al.*, 2014) and estimation of population abundance (Hess *et al.*, 2012). Until recently, it has been very difficult to quantify real dispersal patterns due to the high prevalence of species producing many small larvae with significant planktonic drift times (Strathmann, 1990). Though larval dispersal modelling studies have provided significant insight into probable large-scale patterns of larval connectivity (Cowen *et al.*, 2006), oceanographic model limitations (Nickols *et al.*, 2015), and lack of realistic larval

behaviour (D'Aloia *et al.*, 2015) prevent current dispersal models from being used as a replacement for empirical connectivity estimates. A number of groundbreaking genetic and micro-chemical techniques have recently been developed to experimentally identify the sites of origin of a given set of settling larvae, and thereby quantify empirically larval dispersal (Jones *et al.*, 2005; Kaplan *et al.*, 2010). These studies have revolutionized our thinking with respect to marine larval dispersal, demonstrating levels of self-recruitment (i.e. settlement of larvae at their site of origin) that were previously considered unlikely in marine species with a planktonic larval phase (Jones *et al.*, 2005). Nevertheless, as the number of such studies has grown, an ever-widening variety of self-recruitment levels has been recorded (D'Aloia *et al.*, 2013; Cuif *et al.*, 2015), suggesting that no one paradigm of "open" or "closed" can be applied to all marine species. Clearly much remains to be learned regarding larval dispersal in marine systems. With the advent of high-throughput techniques, such as genotyping-by-sequencing, permitting intensive sampling of species with large population sizes, empirical larval connectivity studies are poised to play a major role in advancing marine science and management.

Despite these advances, experimental studies of larval connectivity remain extremely difficult. The process of "marking" (more precisely defined in the methods) a certain number of larvae emanating from a spawning site and counting up these marked individuals among settlers at a settlement site invariably involves extensive fieldwork and subsequent laboratory analyses to obtain a statistically viable sample. Given the difficulty of performing these studies and the relatively small numbers of "marked" settlers collected, it is important to understand the statistical power of connectivity estimates derived from these data. As we will demonstrate, uncertainty can enter into these studies in numerous ways, but has been relatively ignored in the literature. Though some studies have included specific uncertainty sources in connectivity estimates (Almany *et al.*, 2007; Berumen *et al.*, 2012), no study has taken a global look at the ensemble of uncertainties underlying marine connectivity estimates. This global view is essential to producing robust answers to the major ecological and management questions that these studies have the potential to help resolve.

The goal of this study is to carry out this global examination of uncertainty in marine larval connectivity estimates. We will show that some previously unevaluated, but simple to assess, uncertainty sources are major contributors to the overall uncertainty in connectivity estimates. The ConnMatTools R package, along with sample code detailing the use of the package, is provided to quantify these uncertainties and thereby better integrate future marine larval connectivity studies into marine ecology and management.

## Material and methods

Though all experimental larval connectivity studies are ultimately mark-recapture studies, the details vary as to how larvae are marked and later collected as settlers, and some of these details are important for estimating uncertainty in connectivity. In most cases, larvae are small and numerous and, therefore, one must find a way to identify larvae spawned at a site and settling at another without observing the path taken between the two. Sometimes spatial differences in natural chemical signatures are strong enough that they can be identified in calcifying structures (e.g. otoliths in fish, statoliths in molluscs) of settlers and used to determine their point of origin (Hamilton *et al.*, 2008). In these cases, all individuals spawned at a site are automatically marked with the chemical signature, though accurately separating individuals from different sites will require sufficiently strong chemical gradients between sites.

When natural chemical gradients are insufficient or poorly known, larvae can be artificially marked. One such technique is transgenerational isotope labelling (TRAIL) (Thorrold *et al.*, 2006), in which mature females are injected with a solution enriched in one or more stable isotopes. This artificial chemical signature is transferred to the calcifying structures of larvae produced by injected females, the signal of which is later identified in collected settlers. Due to large population sizes, sampling of mature females is often not exhaustive. As a consequence, the dispersal rate of larvae from one site to another (i.e. the connectivity one hopes to measure) must be inferred from the number of marked individuals collected and the fraction of reproducers marked.

Another technique for assessing connectivity is parentage analysis (Jones *et al.*, 2010), wherein reproducers and settlers are genotyped to uniquely identify (up to a certain statistical power) parent–offspring pairs (POPs) and thereby assess connectivity. Large population size often mandates incomplete genotyping of reproducers, and, therefore, parentage analysis is in some ways conceptually and mathematically similar to TRAIL when applied to marine larval connectivity.

These three techniques (natural chemical markers, artificial chemical markers, and parentage analysis) are currently the most common and reflect well the range of issues associated with the estimation of marine connectivity. All empirical approaches to estimating connectivity generally share a common set of steps that may be subject to uncertainty (Figure 1). We will examine each of these steps, but developing appropriate statistical frameworks is not natural if one follows the experimental order of the steps. Rather, we begin with a subsection on the central inference problem of estimating connectivity from an observed number of marked settlers. Nuances to this basic methodology are developed in subsequent subsections. Specifically, we examine estimation of connectivity when marked settlers cannot be identified with 100% certainty, modifications to our methods when the size of the settler pool is finite, differences between micro-chemical and genetic parentage-based studies in terms of inference regarding the number of "marked" larvae, estimation of production of marked and unmarked eggs for stochastic egg production, and estimation of absolute connectivity values (e.g. local retention) from relative connectivity values. The Methods end with a description of how we demonstrate the statistical frameworks developed by applying them to a sample of available connectivity data.

### Larval dispersal mark-recapture

If we assume that the fraction of "marked" eggs produced at the site of larval origin is known with absolute certainty and marked vs. unmarked settlers can be separated without any doubt (we will reexamine both of these assumptions later on), given collection of $k$ marked settlers out of a total sample of $n$ settlers, then the fraction, $\varphi$, of all settlers at the destination site that originated at the marked site is typically estimated as:

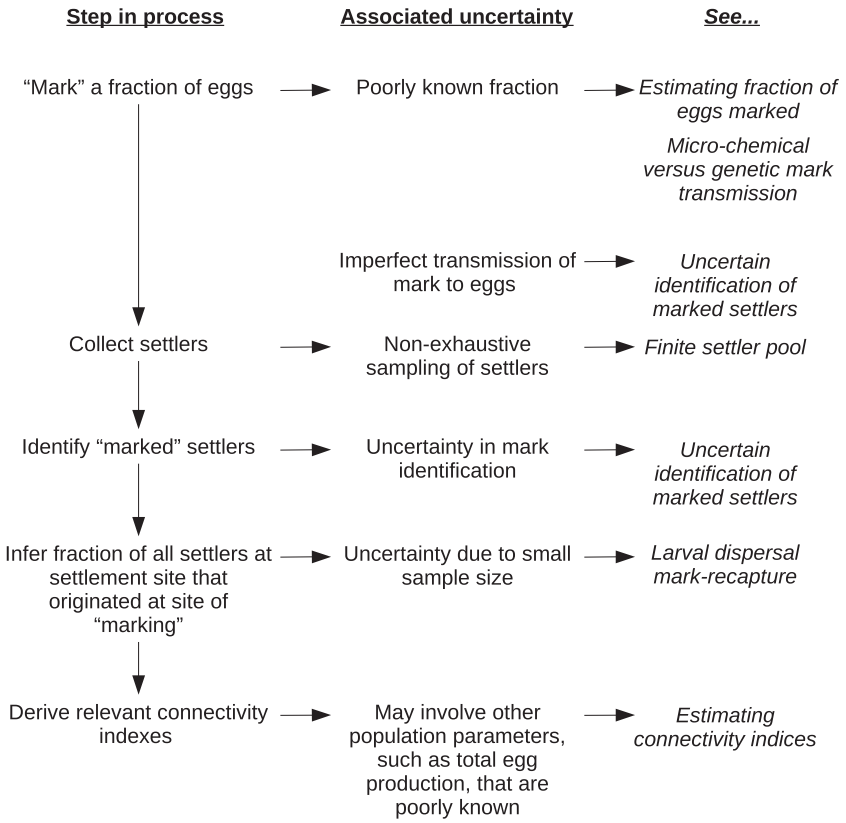


**Figure 1.** Steps in marine larval connectivity empirical studies (left column), their associated uncertainties (middle column) and Methods subsection with pertinent information for assessing each uncertainty (right column).

$$\varphi = \frac{k}{p*n} \tag{1}$$

where $p$ is the fraction of larvae produced at the site of origin that carry the mark. In this equation, $p$ corrects the fraction of marked individuals in the sample for the fact that not all eggs produced at the site of origin carry the mark. Though Equation (1) is often used to evaluate self-recruitment (i.e. when the marked site and the site of settler collection are the same), it is general and can be used to evaluate relative (to total settlement at a site) connectivity between different sites.

Despite Equation (1) being widely used to calculate connectivity rates, it is problematic for a couple of reasons. For one, if $p$ is smaller than $k/n$, then the fraction will be bigger than 1, which is impossible. This has not been formally recognized in the literature, but it is conceptually possible if this equation is applied carelessly. Second, this equation assumes that the sample of settlers is representative of the wider pool of settlers at the settlement site, which may not be the case, particularly if sample size is small. For example, if only 1 marked individual is found out of 10 settlers, one could wonder how likely it is that the true fraction of marked individuals in the entire pool of settlers differs significantly from 10%. Anyone who has tossed a coin multiple times can attest to the fact that it sometimes takes many flips for the sample average number of heads to approach the population mean of 50%. The same is true for collecting settlers at a site.

Both of these issues can be addressed using standard approaches from mark-recapture/site-occupancy studies (MacKenzie *et al.*, 2005). Marked individuals are found among a sample of settlers collected at a destination site via two approximately binomial processes (Figure 2): first, the sample of settlers is randomly selected among the pool of settlers, an unknown fraction, φ, of which originated at the marked site; second, the subset of collected settlers that originated at the marked site are derived from a random draw from the pool of eggs produced at the spawning site, of which a known fraction, $p$, are marked (assuming that marked and unmarked eggs are well mixed). Therefore, the likelihood function for φ given $k$ marked settlers out of a sample of $n$ settlers is proportional to a binomial distribution with probability $p\varphi$:

$$\begin{aligned} \mathrm{L}(\varphi|n,k,p) &\propto \Pr(k|n,p\varphi) = \mathrm{Binom}(k|n,p\varphi) \\ &= \binom{n}{k}(p\varphi)^k(1-p\varphi)^{n-k} \end{aligned} \tag{2}$$

By way of Bayes' Theorem, we can convert this likelihood into a (posterior) probability density function for φ:

$$\Pr(\varphi|n,k,p) = \frac{\Pr(k|n,p\varphi)\Pr(\varphi)}{\int_0^1 \Pr(k|n,p\vartheta)\Pr(\vartheta)\,d\vartheta} \tag{3}$$

where $\Pr(\varphi)$ is the prior distribution for φ. As one generally has little prior information on the value of φ, a "non-informative" prior is most appropriate. Non-informative priors for binomial models typically take the form of symmetric Beta distributions: $\Pr(\varphi) = f(\varphi;\alpha,\beta) \propto \varphi^{\alpha-1}(1-\varphi)^{\beta-1}$, where $\alpha = \beta = 1$ for a

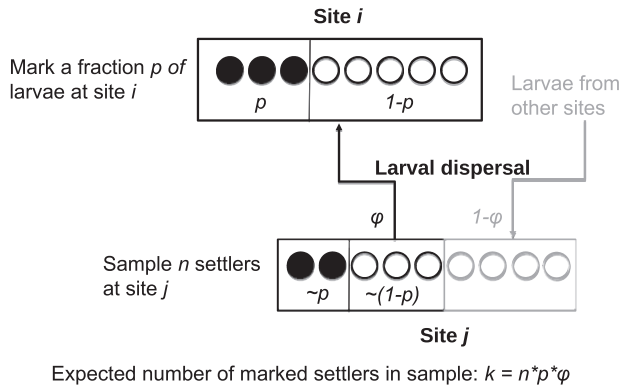


**Figure 2.** Conceptual diagram of binomial processes determining observed number of marked individuals in a given sample of settlers.

uniform prior, and $\alpha = \beta = 1/2$ for the Jeffreys or Reference prior (Berger *et al.*, 2015). If a uniform prior is used, the posterior distribution for φ is closely related to the Beta distribution and can be evaluated analytically:

$$\Pr(\varphi|n,k,p) = \frac{(p\varphi)^k(1-p\varphi)^{n-k}}{\frac{1}{p}\int_0^p x^k(1-x)^{n-k}dx} = \frac{f(p\varphi;k+1,n-k+1)}{\frac{1}{p}\int_0^p f(x;k+1,n-k+1)dx} \quad (4)$$

Given the probability density functions in Equations (3) and (4), one can calculate CIs and most probable values for φ. We will refer to the uncertainty in connectivity estimates derived from these equations as the "small sample size uncertainty" as it is closely related to the potential for biased sampling of the settler pool due to a limited sample size.

Equations (2)–(4) are a special application of the general Cormack-Jolly-Seber (CJS) mark-recapture model (Schwarz, 2001) to the case of a single recapture event. In the context of CJS models, $p$ and φ are generally referred to as the "detection probability" and "apparent survival", respectively. Unlike many applications of CJS models, having only a single recapture event does not allow one to independently estimate $p$. Rather, $p$ must be separately estimated and included as a fixed parameter in the model. Uncertainty in $p$ can, however, be included by, e.g. ensemble averaging over possible values of $p$.

The logic applied above can be naturally extended to the case of multiple different types of marked settlers, e.g. if adults in distinct source sites are marked with different artificial chemical signatures. In this case, the binomial probability distribution in Equationss (2)–(4) is replaced by a multinomial distribution:

$$\Pr(\varphi_1 \ldots \varphi_s|n,k_1 \ldots k_m,p_1 \ldots p_s) \propto \left(1-\sum p_i\varphi_i\right)^{n-\sum k_i}\prod(p_i\varphi_i)^{k_i}\Pr(\varphi_1 \ldots \varphi_s) \quad (5)$$

Appropriate non-informative priors for such models generally take the form of Dirichlet distributions with uniform parameters (Berger *et al.*, 2015).

## Uncertain identification of marked settlers

The connectivity estimation procedure described above can be extended to the case where marked-unmarked status of collected settlers is not known with absolute certainty. If $\rho_j$ is the probability that the $j$th collected settler is marked, then the likelihood function for φ is:

$$\mathrm{L}(\varphi|p,\rho_1 \ldots \rho_n) \propto \Pr(\rho_1 \ldots \rho_n|p\varphi) = \prod_{j=1}^{n}[(1-p\varphi)(1-\rho_j)+p\varphi\rho_j] \quad (6)$$

Equation (6) can be used to develop a Bayesian estimator for the probability density function for φ following the same logic as in Equation (3).

One particularly common case where Equation (6) is applicable is when a "score" (i.e. a numerical value related to the likelihood of being marked) is generated for each collected settler. For micro-chemical studies, the score is typically an isotope concentration or ratio from the core of the settler's calcifying structure, whereas for parentage analysis, the score is often a likelihood that a settler was not produced by random combination of non-genotyped individuals (i.e. the "log of the odds ratio" (LOD); Gerber *et al.*, 2003). If normalized probability densities for the theoretical distributions of scores among marked and unmarked individuals can be estimated, e.g. from laboratory marked and unmarked individuals for TRAIL, or simulations based on observed allelic frequencies for parentage analysis (Gerber *et al.*, 2003), then:

$$\rho_j = \frac{P_{\mathrm{M}}(x_j)}{P_{\mathrm{M}}(x_j)+P_U(x_j)} \quad (7)$$

and:

$$\mathrm{L}(\varphi|p,x_1 \ldots x_n) \propto \Pr(x_1 \ldots x_n|p\varphi) \propto \prod_{j=1}^{n}[(1-p\varphi)P_U(x_j)+p\varphi P_M(x_j)] \quad (8)$$

where $x_j$ is the score of the $j$'th sampled settler, and $P_{\mathrm{M}}(x)$ and $P_{\mathrm{U}}(x)$ are the probability density functions for scores of marked and unmarked individuals, respectively. Estimates for φ based on Equation (8) will include uncertainty due to both sample size and the potential for false identification of marked and unmarked individuals. In particular, estimates for φ can be used to estimate the probability that an individual settler is marked correcting for prevalence of marked individuals in the sample and the potential for false assignment:

$$\tilde{\rho}_j = \frac{p\varphi P_{\mathrm{M}}(x_i)}{(1-p\varphi)P_{\mathrm{U}}(x_i)+p\varphi P_{\mathrm{M}}(x_i)} \quad (9)$$

Equations (7)–(9) provide a straightforward approach to assessing probability of an individual settler being marked when scores are generated for each settler that is applicable to both TRAIL and parentage-based connectivity studies. When used in the context of parentage studies, these equations provide an alternative methodology for identifying POPs that is conceptually similar to the approach to parentage analysis proposed by Christie *et al.* (2013).

## Finite settler pool

The probability distribution in Equations (2)–(4) is only exactly valid in the limit of a well-mixed pool of settlers of *infinite* size. However, in reality the settler pool is finite and the probability

distribution may need to be adjusted to take this into account. To see this, consider the case where $p = 1$ (i.e. all eggs produced at the marked site are marked) and all settlers to a site are collected. In this case, there is no uncertainty in φ (it is simply $k/n$), but Equation (3) would not predict this. This is due to the fact that the probability of sampling a marked individual changes as one removes individuals from a finite settler pool. Equations (2)–(4) can be corrected to take this into account using hypergeometric distributions, which are similar to binomials, except that true–false values are sampled without replacement from a finite pool of values, a known fraction of which are true. For example, if we know there are a total of $K$ marked individuals in the settler pool of size $N$, then the probability of observing $k$ marked individuals in a sample of size $n$ is:

$$\text{Hyper}\,(k|n,N,K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{10}$$

The probability of $K$ marked settlers in the settler pool is binomially distributed:

$$\text{Binom}\left(K|\tilde{K},p\right) = \binom{\tilde{K}}{K}p^{\tilde{K}}(1-p)^{\tilde{K}-K} \tag{11}$$

where $\tilde{K}$ is the total number of individuals originating at the source site in the settler pool at the destination site (a subset $K$ of which are marked). Combining these and summing over all possible values of $K$ leads to the probability mass function for $k$, which can then be inverted using Bayes' Theorem to obtain the probability mass function for $\tilde{K}$:

$$\Pr\left(\tilde{K}|k,n,N,p\right) = \frac{\Pr\left(\tilde{K}\right)\sum_{K=k}^{\tilde{K}}\text{Binom}\left(K|\tilde{K},p\right)*\text{Hyper}\,(k|n,N,K)}{\sum_{L=k}^{L=N}\Pr\left(L\right)\sum_{K=k}^{\tilde{L}}\text{Binom}\,(K|L,p)*\text{Hyper}\,(k|n,N,K)} \tag{12}$$

where $Pr\left(\tilde{K}\right)$ is the prior distribution for $\tilde{K}$, generally taken to be a constant for a non-informative prior. Given $\tilde{K}$, relative connectivity $\varphi = \tilde{K}/N$.

## Micro-chemical vs. genetic mark transmission

How the fraction of marked eggs, $p$, relates to the number of adults "marked" in a larval connectivity study is one potential important difference between genetic and micro-chemical approaches to measuring dispersal. In techniques that use micro-chemical signatures (natural or artificial), mark transmission can only pass through females. Therefore, assuming all females produce the same number of eggs (a dubious assumption to be revisited), $p = p_f$, the fraction of the mature female population that was marked.

By contrast, nuclear genetic material is transmitted by both males and females, and, therefore, for parentage studies, the fraction of eggs produced at the "marked site" for which at least one of the parents was genotyped may also depend on the fraction of the adult male population that is genotyped, $p_\text{m}$. If probability of two individuals mating is independent of whether or not one or both has been genotyped (i.e. genotyped and non-genotyped adults are well mixed at the site of egg production) and all individuals contribute equally to the number of eggs produced, then the fraction of eggs with at least one genotyped parent is given by (Harrison *et al.*, 2012):

$$p = 1-(1-p_\text{f})(1-p_\text{m}) = p_\text{f}+p_\text{m}-p_\text{f}p_\text{m} \tag{13}$$

Using Equation (13), the probability distribution in Equation (3) can be applied to this more complicated case of maternal and paternal mark transmission. However, one can go further and test the validity of the assumption of equal probability of mating by comparing the number of settlers collected with one male, one female or two known parents. Again assuming equal probability of mating, one would predict the fraction of marked eggs with one known parent to be:

$$F_i = \frac{p_i(1-p_j)}{1-(1-p_\text{f})(1-p_\text{m})} \tag{14}$$

where $\{i,j=\{f,m\}$ $(\{m,f\})$ for a known female (male) parent. Similarly, the fraction of marked eggs with two known parents would be:

$$F_\text{mf} = \frac{p_\text{m}p_\text{f}}{1-(1-p_\text{f})(1-p_\text{m})} \tag{15}$$

This comparison could be qualitative, or quantitative if sufficient POPs are identified to permit inference of probability distributions for $p_\text{m}$ and $p_\text{f}$:

$$\Pr\left(p_\text{m},p_\text{f}|k_\text{m},k_\text{f},k_\text{mf}\right) \propto \Pr(p_\text{m},p_\text{f})\prod_{i=m,f,mf}F_i^{k_i} \tag{16}$$

where $k_\text{m}$, $k_\text{f}$, and $k_\text{mf}$ are the number of settlers with known male, female, and two-known parents in the sample and P r$(p_\text{m},p_\text{f})$ is the prior distribution for $p_\text{m}$ and $p_\text{f}$. One could compare the resulting posterior distributions for $p_\text{m}$ and $p_\text{f}$ with independent experimental estimates of $p_\text{m}$ and $p_\text{f}$ (e.g. based on the number of genotyped adults and an estimate of the total adult population size) to assess the validity of the uniform probability of mating hypothesis. Note that in principle simultaneous estimation of φ, $p_\text{m}$, and $p_\text{f}$ is possible by substituting Equation (13) in the likelihood in Equation (2), but this would require many POPs and certainty that genotyped and non-genotyped reproducers are well mixed.

## Estimating fraction of eggs marked

Though numerous studies have considered uncertainty in the adult population size at the site of marking when calculating $p$, the fraction of marked eggs, uncertainty in egg production itself has rarely been taken into account. It is well know that egg production in marine organisms varies as a function of age and size (Gunderson, 1997). Furthermore, reproduction itself is a stochastic process in many species, with not all individuals reproducing at every mating opportunity. In some cases, there may be enough information to correct for these processes when estimating the fraction of marked eggs from the fractions of marked mature males and females. For example, if reproduction is reasonably synchronous and universal among mature individuals *and* size information is available for *both* marked and unmarked individuals *and* egg production as a function of size is known and reasonably deterministic, then one could simply calculate total marked and unmarked egg production from the numbers and size-frequency of marked and unmarked individuals.

More often, precise estimation of the fraction of marked eggs is impossible because insufficient information is available on the size structure of marked and (more problematic) unmarked individuals, and/or some of the processes underlying reproduction are fundamentally stochastic. However, the magnitude of the uncertainty in the fraction of marked eggs can be estimated if the magnitude of variation in individual reproductive output is known. For the case of maternal mark transmission, if $w$ mature females are marked out of a total population of $W$ mature females, then the fraction of marked eggs can be estimated using:

$$p = \frac{\sum_{i=1}^{w} e_i}{\sum_{i=1}^{W} e_i} = \frac{\sum_{i=1}^{w} e_i}{\sum_{i=1}^{w} e_i + \sum_{i=w+1}^{W} e_i} \tag{17}$$

where $e_i$ is the egg production of individual $i$, which is assumed to be a random variable drawn from a distribution with known mean $\bar{e}$ and SD $\sigma_e$. If $w$ and $W - w$ are sufficiently large, then by the central limit theorem this can be approximated by:

$$p \sim \frac{X}{X + Y} \tag{18}$$

where $X$ and $Y$ are normally-distributed (or gamma-distributed to avoid negative numbers) random variables satisfying:

$$\begin{aligned} X &\sim \mathcal{N}(w*\bar{e}, w*{\sigma_e}^2) \\ Y &\sim \mathcal{N}((W-w)*\bar{e}, (W-w)*{\sigma_e}^2) \end{aligned} \tag{19}$$

Though in principle the distribution for $p$ could be calculated analytically as the convolution of the distributions in Equation (19) constrained by Equation (18), the result is quite complex. It is easier to estimate the distribution for $p$ numerically by drawing random values from the distributions in Equation (19), and to incorporate these values for $p$ into the overall uncertainty in connectivity using a bootstrap approach (i.e. average the probability distributions for φ from each potential value of $p$). Often, there is also uncertainty in the total number of mature females, $W$, and this uncertainty can be integrated into overall uncertainty in $p$ by bootstrapping over potential values for $W$.

Note that in some cases $w$ and $W$ may not represent the number of individual fish, but rather the number of groups of fish, each of which reproduces as a single unit. One example of this would be colonial species that reproduce synchronously.

Equations (17)–(19) may be extended to the case of maternal and paternal mark transmission by using Equation (17) as a replacement for $p_f$ in, e.g. Equation (13) and repeatedly randomly sampling the distributions of individual egg production to estimate the distribution of $p$. Whether or not a similar replacement for $p_m$ also needs to be included will depend on reproductive behaviour. One would expect that total sperm production would be irrelevant for species that spawn in pairs as sperm are generally not limiting, but may be important for species that spawn in aggregation and for which males are in direct competition with one another for fertilizing eggs.

## Estimating connectivity indices

We have so far treated connectivity as being synonymous with "relative connectivity", i.e. the fraction of all settlers at a site that originated at some specific site. When the larval production and settlement study sites are the same, this is known as the self-recruitment. This measure of connectivity is useful for assessing the openness of populations, but it is only related to population persistence under certain specific conditions regarding population stability and homogeneity, and even then provides only an assessment of relative persistence (Lett *et al.*, 2015). Though the majority of empirical larval connectivity studies have focused on self-recruitment or relative connectivity, population persistence is best assessed by estimating the connectivity matrix, the elements of which indicate the settlement rate from one site to another relative to total egg production at the spawning site (Burgess *et al.*, 2014); the diagonal elements of connectivity matrix are typically referred to as the "local retention"). Elements of the connectivity matrix can be estimated from the fraction of settlers originating at the marked site, φ, if one has an estimate of the total number of settlers at the destination site and the total egg production at the site of origin:

$$c = \frac{S*\varphi}{\sum_{i=1}^{W} e_i} \sim \frac{S*\varphi}{X + Y} \tag{20}$$

where $c$ is the element of the connectivity matrix corresponding to larval transport from the spawning site where individuals are marked to the site where settlers are collected, $S$ is the total settlement to the destination site, the $e_i$ are the egg productions of the $W$ mature individuals at the spawning site, and $X + Y$ are as in Equations (18) and (19).

Equation (20) introduces new potential sources of uncertainty in connectivity estimates. s total egg production appears in Equations (17) and (18) for φ and Equation (20) for the connectivity matrix, uncertainty in φ and $X + Y$ must be jointly estimated (e.g. using a bootstrap procedure over possible values of $X$ and $Y$). Uncertainty in total settlement at the destination site, $S$, will depend on the experimental protocol used, but is typically independent of the uncertainty in $\varphi/X + Y$. As such, it can be estimated separately and then combined with uncertainty in $\varphi/X + Y$.

## Application of methods to results from existing literature

The majority of the methods described above are implemented in the ConnMatTools R package (see Supplementary Materials S1, S2 and S3). We used this package to assess the magnitude of uncertainty in a non-exhaustive subsample of published larval connectivity estimates. The studies considered (Table 1) include micro-chemical and genetic parentage-analysis studies, as well as examples of exhaustive and non-exhaustive marking and recapture protocols. From each study, the fraction of eggs that were marked, settler sample size, number of marked settlers observed in the sample and any information available on exhaustivity of adult and settler sampling were extracted and used to estimate small sample size uncertainty in connectivity estimates, as well as uncertainty related to inaccuracy in the estimate of the fraction of eggs marked (where possible). For brevity, uncertainties for only a subset of connectivity measurements are estimated for studies including multiple connectivity measurements (e.g. different time periods or sites). For each connectivity measurement, the 95% CI was evaluated based on Equation (4) and, where appropriate, Equation (4) was ensemble averaged over possible values of the fraction of marked eggs, $p$, assuming a uniform distribution of

**Table 1.** Application of small sample size uncertainty model to a subset of published larval connectivity results.

| Study (Study type) | Measurement[a] | Fraction of eggs marked, *p* *Mode [min; max of CI]* | Marked settlers collected | Settler sample size | Exhaustive?[b] | Study connectivity estimate *Mode [min; max of CI]* | Estimate including small sample size (Equation 4)[c] *Mode [min; max of CI]* | Estimate including uncertainty due to sample size and fraction eggs marked[c,d] *Mode [min; max of CI]* |
|---|---|---|---|---|---|---|---|---|
| Jones *et al.* (1999) **(Tetracycline mark)** | SR | [0.005; 0.02] | 15 | 5000 | ? | [0.15; 0.6] | 0.24 [0.15; 0.40][e] | 0.19 [0.12; 0.67] |
| Jones *et al.* (2005) **(Tetracycline Mark & Parentage)** | SR, 2002 | 1 | 10 | 63 | Yes | 0.16 | – | – |
| Almany *et al.* (2007) **(TRAIL)** | SR, A. percula | 1 | 9 | 15 | ? | 0.60 | 0.60 [0.35; 0.80] | – |
| | SR, C. vagabundus | 0.173 [0.144; 0.200] | 8 | 77 | ? | 0.60 [0.52; 0.72] | 0.60 [0.31; 0.96] | 0.60 [0.31; 0.96] |
| Planes *et al.* (2009) **(Parentage)** | SR, 2004 | 1 | 56 | 133 | ? | 0.42 | 0.42 [0.34; 0.51] | – |
| | Connectivity, Restorf Island | 1 | 5 | 50 | ? | 0.10 | 0.10 [0.044; 0.21] | – |
| Christie *et al.* (2010) **(Parentage)** | SR to Big Island | 0.0006 [0.0002; 0.002] | 4 | 566 | ? | 12 [3.5; 35][f] | 1 [0.54; 1] | 1 [0.52; 1] |
| | Connectivity, Miloli'i to Ho'okena | 0.0007 [0.0004; 0.0024] | 1 | 68 | ? | 20 [6.1; 34] | 1 [0.22; 1] | 1 [0.22; 1] |
| Berumen *et al.* (2012) **(Parentage)** | SR, A. percula | 0.95 | 103 | 161 | ? | 0.64 | 0.67 [0.59; 0.75] | – |
| | SR, C. vagabundus | 0.221 [0.187; 0.271] | 9 | 103 | ? | 0.38 [0.32; 0.47] | 0.40 [0.21; 0.71] | 0.38 [0.20; 0.72] |
| D'Aloia *et al.* (2013) **(Parentage)** | SR | 1 | 9 | 194 | Yes | 0.046 | – | – |
| Schunter *et al.* (2014) **(Parentage)** | SR | 1 | 25 | 382 | ? | 0.065 | 0.070[0.045; 0.095] | – |
| Cuif *et al.* (2015) **(TRAIL)** | SR, entire study | 0.194 [0.160; 0.230] [d] | 13 | 564 | No | – | 0.12 [0.070; 0.20] | 0.12 [0.069; 0.21] |
| | SR, March 2012 | 0.197 [0.164; 0.232] [d] | 11 | 82 | No | – | 0.68 [0.39; 0.97] | 0.68 [0.38; 0.97] |

[a]"SR" = self-recruitment.

[b]Column indicates whether or not collection of settlers was exhaustive. "?" = unknown if settler collection was exhaustive. In many cases, it was not immediately clear to what extent settler collection was exhaustive, in which case we made the conservative assumption that sample size was much smaller than the total settler pool.

[c]For parentage studies, it was generally not possible to determine if a correction for dual mark transmission was appropriate or needed, so the values of *p* in the published article were used directly.

[d]A uniform distribution of probability for the value of *p* between the lower and upper bounds indicated in the relevant study was assumed when calculating an ensemble average probability distribution for relative connectivity, except for Cuif *et al.* (2015), for which Equations (17)–(19) were used to estimate uncertainty in *p* assuming synchronous reproduction on each of the 508 branching coral colonies, 100 of which are marked, and an average of 1 unit of reproduction per colony with a SD equal to the mean (i.e. also 1). The assumption of SD equal to the mean is likely appropriate for the monthly connectivity estimate, but overly conservative (i.e. tending towards larger uncertainty) for the connectivity assessment for the entire study (being an average over multiple reproductive opportunities).

[e]Based on mean value for *p*.

[f]Calculated based on values in study, but not presented in the original study itself.

**Table 2.** Local retention estimates integrating multiple different uncertainty sources for settlement of humbug damselfish to focal reef in southwest lagoon of New Caledonia in March 2012[a]

| Uncertainty level | Local retention estimates[b] *Median [min; max of 95% CI]* |
|---|---|
| No uncertainty | 0.41 |
| Fraction of marked eggs[c] | 0.41 [0.35; 0.49] |
| Small sample size | 0.41 [0.23; 0.58] |
| Frac. marked eggs[c] and Sample size | 0.41 [0.23; 0.58] |
| and Egg production[c] | 0.41 [0.23; 0.59] |
| and Settler pool size[d] | 0.41 [0.22; 0.64] |

[a]From Cuif *et al.* (2015). 19.7% of colonies marked. 11 of 82 collected settlers identified as marked.
[b]Units = [settlers][breeding colony]$^{-1}$[month]$^{-1}$.
[c]Equations (17)–(20) were used to estimate uncertainty in total egg production and fraction of eggs marked, *p*, assuming synchronous reproduction on each of 508 branching coral colonies, 100 of which are marked, and an average of 1 unit of reproduction per colony with an SD equal to the mean (i.e. also 1).
[d]The total size of the settler pool was modelled following Cuif *et al.* (2015) as a Gamma distributed random variable with mean of 305 and SD of 36.7.

probability for $p$ between bounds provided in the original article (there was generally not enough information to justify using a more complex distribution).

Larval connectivity data collected by the authors for the humbug damselfish (*Dascyllus aruanus*) of New Caledonia was also used to examine the impact of different uncertainty sources that could not be examined using the published literature due to a lack of necessary data. Over the ~3-month reproductive season of this species and for 2 consecutive years (2011 and 2012), all adult damselfish on 100 branching coral colonies out of a total of 508 colonies at a site in the centre of the southwest lagoon of New Caledonia were repeatedly injected with a solution enriched in $Ba^{137}$. Beginning one month after initial injection, ~100 settlers per month were collected at this same site to identify marked individuals. Details of the study mark and recapture protocols are available in (Cuif *et al.*, 2015).

Uncertainty estimates for self-recruitment, *as well as local retention*, were generated for these data from New Caledonia. The number of marked and unmarked damselfish colonies at the New Caledonia study site is known with certainty, but the reproductive output of each colony is uncertain. As an illustration of the possible effects of uncertainty in reproductive output and absent concrete data on variability in egg production, we estimated uncertainty in $p$ based on 1000 random draws from Equations (17)–(19) assuming that each of the 508 colonies at the study site produced on average 1 unit of reproductive output with a SD equal to the mean (i.e. SD = 1). Cumulative uncertainty in local retention due to small sample size, fraction of eggs marked and total egg production was estimated using a bootstrap approach, randomly drawing 1000 potential values for relative connectivity from the probability distribution in Equation (4) for each of the 1000 previously calculated potential values for $p$ and total reproductive output. Uncertainty due to the total number of settlers was similarly integrated, randomly drawing a total settlement value from a Gamma distribution with mean 305 and SD 36.7 (Cuif *et al.*, 2015) for each potential value of total egg production and relative connectivity.

Genotype data from damselfish in New Caledonia were used to demonstrate the approach to parentage analysis described in Equations (6)–(9). Once per reproductive season, genetic material was obtained from adults on the same 100 branching coral colonies where TRAIL was used. Adults and larvae later collected as part of the TRAIL study were genotyped at 17 microsatellite loci. LOD values were calculated for putative POPs, as well as for simulated true and false POPs based on allelic frequencies observed in adult individuals, using the FAMOZ software package (Gerber *et al.*, 2003). Observed and simulated LOD values were then compared to estimate the probability distribution for the number of real POPs in the sample of recruits following Equations (8) and (9). As this dataset has yet to be fully examined and published, results from a non-random subsample of 200 recruits are presented here purely to demonstrate the approach.

All connectivity uncertainty calculations (i.e. Table 2 and last two columns of Table 1) were generated assuming a uniform prior probability distribution for relative connectivity, φ. Using a uniform prior was simpler and assured alignment of most probable values for φ with what one would naively expect from Equation (1). Nevertheless, methods for using both uniform and non-uniform priors are implemented in the ConnMatTools R package. Differences between results for uniform and non-uniform priors were typically small.

## Results

Applying Equation (4) to typical values for settler sample size and number of marked individuals in the sample demonstrates that small sample size is often a major source of uncertainty (Figure 3). For a sample size of 20 settlers (a low value, but not unheard of for a single site and settlement period; e.g. Saenz-Agudelo *et al.*, 2011) and $p = 0.5$, the widths of the 95% CI for relative connectivity φ are 0.27, 0.68, and 0.40 for most probable values of φ of 0, 0.5 and 1, respectively (Figure 3a). If sample size is increased to 200 individuals, 95% CI widths reduce to 0.03, 0.24, and 0.14, respectively (Figure 3b). Overall, 95% CI width decreases as the square root of settler sample size unless the most probable value for relative connectivity is zero, in which case it decreases linearly (Figure 4). Note that, even if Equation (1) predicts $\varphi = 0$ or $\varphi \geq 1$, the probability density function is by definition constrained to the interval $[0, 1]$, and will be peaked at 0 or 1 in these cases (Figure 3).

Sampling a large fraction of the (in reality finite) settler pool can reduce small sample size uncertainty, but only if sampling is close to exhaustive and the fraction of eggs marked, $p$, is also nearly 100% (Figure 5). For either a total settler pool 20% greater than the sample size (Figure 5, second dot from left on red curve) or 90% of eggs marked (Figure 5, first dot from left on green curve), the 95% CI width is ~40% of that for an infinite settler pool. For a settler pool twice the sample size, 95% CI width is ~60–70% that of an infinite settler pool ($p \geq 0.9$). Absolute size of the sample did not affect this relative result (results not shown), though larger samples do have smaller levels of absolute uncertainty. Results for an infinite settler pool are, therefore, a reasonable and conservative estimate of uncertainty except if marking and collection of eggs and settlers, respectively, are close to exhaustive (e.g. both approximately >80%).

The fraction of eggs carrying the "mark" of a known parent can differ significantly depending on whether mark transmission from parent to offspring is maternal (e.g. in TRAIL) or dual maternal and paternal (e.g. as typically assumed in parentage

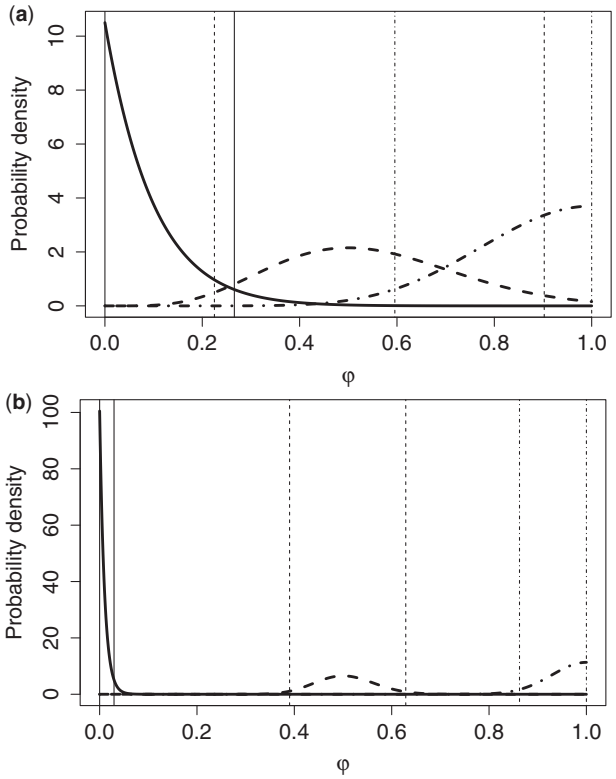


**Figure 3.** Probability distribution for relative connectivity, $\varphi$, due to small sample size uncertainty for settler samples of 20 (a) and 200 (b) individuals. The fraction of eggs marked $p$=0.5. Thick solid, dashed and dot-dashed continuous curves are for 0%, 25% and 50% of collected settlers being marked individuals (corresponding to most probable values of relative connectivity of 0, 0.5 and 1, respectively). The vertical solid, dashed and dot-dashed lines indicate the bounds of the 95% CI for the curve with the corresponding line style.

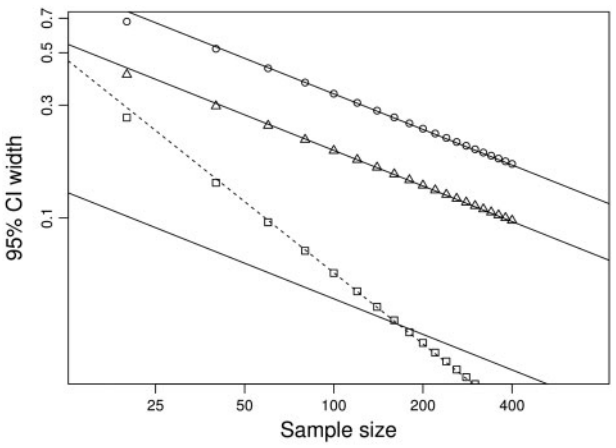


**Figure 4.** Log-log plot of settler sample size vs. 95% CI width for relative connectivity based on Equation (4). Squares, circles and triangles are results for most probable relative connectivity values of 0, 0.5, and 1, respectively. The fraction of eggs marked $p = 0.5$. Solid lines are best log-linear fit with a slope of $-1/2$ to data for each relative connectivity value. The dashed line is best log-linear fit with a slope of $-1$ to squares (relative connectivity of 0).

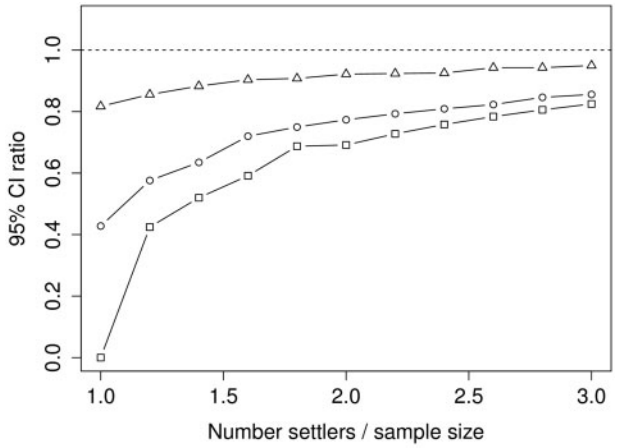


**Figure 5.** The ratio of the 95% CI width for finite and infinite settler pools as a function of the ratio between the total size of the settler pool and the settler sample size. Curves with squares, circles and triangles are for fraction of eggs marked $p = 1.0$, 0.9, and 0.5, respectively. The dashed horizontal line indicates a CI ratio of 1, meaning finite and infinite settler pools lead to equivalent uncertainties. The settler sample size was fixed at 200 individuals, though results were very similar for other sample sizes. The number of marked settlers found in the sample was adjusted so that the most probable value for relative connectivity was 50% in all cases.

studies). Assuming that a small, but equal, fractions of males and females are marked (e.g. genotyped), then the fraction of eggs marked, $p$, is twice as large for dual mark transmission as it is for pure maternal mark transmission (red curve in Figure 6). This difference decreases linearly to zero as the fraction of adults marked increases to 100%.

## Uncertainty in published relative connectivity estimates

The importance of Equation (4) for relative connectivity estimates in the literature varies considerably as a function of study design (Table 1). In some cases, marking of reproducers and collection of settlers are both exhaustive and there is no uncertainty in connectivity estimates associated with sample size (Jones *et al.*, 2005; D'Aloia *et al.*, 2013). In other cases, uncertainty related to sample size can be considerable even if all eggs are marked. For example, in the groundbreaking TRAIL article of Almany *et al.* (2007), uncertainty in self-recruitment due to small sample size is of the same order of magnitude for the two species examined despite one being exhaustively marked and the other not (95% CI width of 0.45 vs. 0.65, respectively) due to the smaller sample size of the prior than the latter (15 vs. 77, respectively). Furthermore, uncertainty in self-recruitment due to sample size is considerably larger than that induced by uncertainty in $p$ (as estimated by Almany *et al.*, 2007), and including both uncertainties only marginally changes the result based on small sample size uncertainty alone. This is also true for several other studies we examined, though uncertainty in $p$ eventually becomes dominant given sufficiently large sample size (e.g. Jones *et al.*, 1999). It is important to note that it was not clear in many studies if sampling of settlers was exhaustive or not.

Results from Christie *et al.* (2010) are unique among studies considered in that predicted connectivity based on Equation (1) is >1 (e.g. 11.8 or 1180% self-recruitment back to Big Island of Hawaii). Using Equation (4) will constrain connectivity to be ≤1,

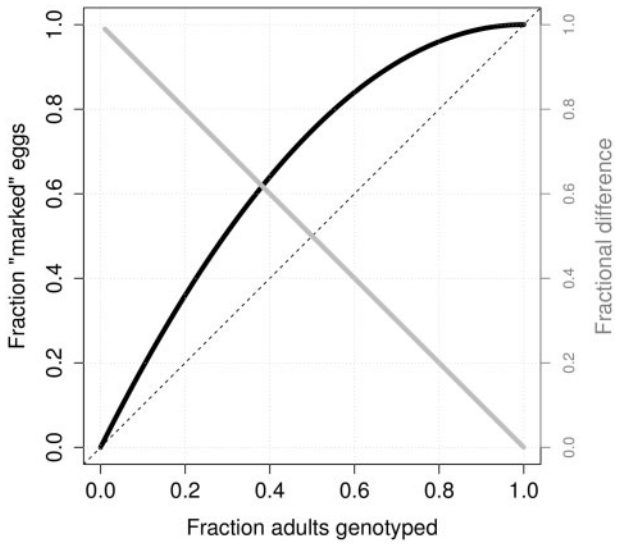


**Figure 6.** The fraction of eggs derived from at least one genotyped parent (p) as a function of the fraction of all adults genotyped, assuming equal numbers of males and females genotyped and equal probability of mating. The dashed diagonal line is the 1–1 line that would occur if only females contributed genetic material (i.e. single mark transmission). The gray line (descending from upper-left to lower-right of graph) gives the fractional difference in dual versus single mark transmission estimates for $p$ (i.e. a value of 0 indicates they are the same, 1 indicates dual mark transmission has a $p$ that is 100% more than that for single mark transmission).

but blind application of the method to this case is unwise. The probability of observing a sample for which Equation (1) predicts 1180% self-recruitment when the true self-recruitment is necessarily ≤100% is exceedingly small (≤0.04% for 4 POPs identified out of 566 settlers and $p = 0.0006$). This suggests that this data violate some of the assumptions underlying connectivity estimation based on Equations (1)–(4). The authors suggest sweepstake reproduction, i.e. that genotyped adults are much more successful reproducers than the average adult (which would increase $p$), but this would require consistent and significant sampling bias towards the most successful reproducers across multiple distinct sites. Other possibilities include false POPs and overestimation of the adult population size. False POPs seem unlikely as the authors genotyped five additional loci for putative POPs and found perfect agreement. Overestimation of the adult population size is a possibility, though this would require the true adult population size to be roughly 11.8 times smaller than experimentally estimated to achieve self-recruitment ≤100% (5.9 times smaller if dual mark transmission, i.e. Equation (13), is appropriate, but was not taken into account).

## Uncertainty in local retention estimates

Local retention estimates for TRAIL data from March 2012 for the humbug damselfish of New Caledonia demonstrate integration of multiple uncertainty sources into connectivity estimates. As with uncertainties in published relative connectivity estimates, uncertainty in local retention was dominated by uncertainty due to the small settler sample size (unless true variability in egg production between damselfish colonies is much larger than was assumed), though uncertainty in the size of the total settler pool

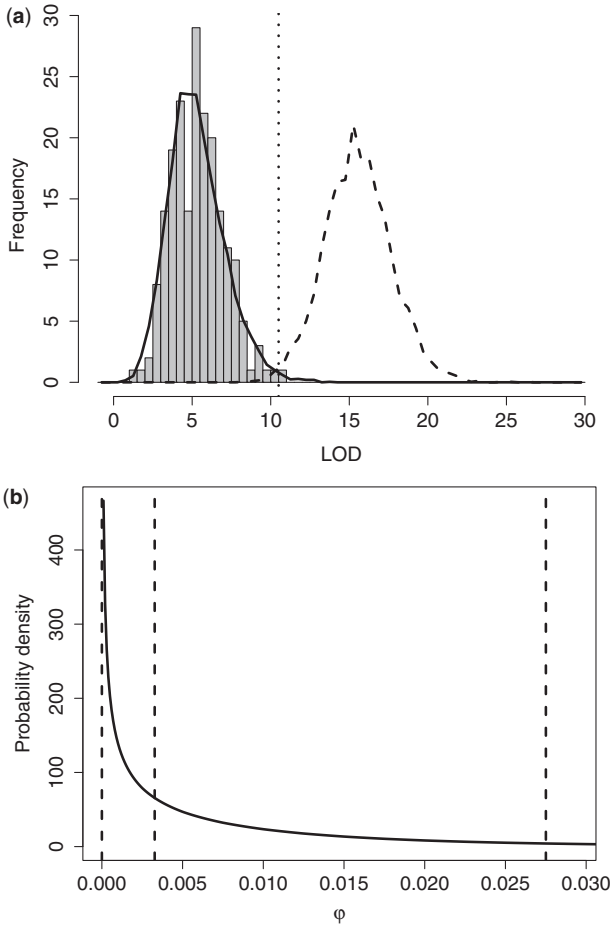


**Figure 7.** Connectivity estimation based on comparison of observed distribution of LODs to predicted distributions for true and false POPs. In (a), gray bars indicate observed distribution of maximum LOD values for each sampled recruit, the thick solid curve indicates predicted LOD distribution for false POPs, and the thick dashed curve indicates predicted LOD distribution for true POPs. The vertical dotted line is the LOD threshold that balances type I and type II errors for samples including equal numbers of true and false POPs. The predicted probability distribution for relative connectivity, $\varphi$, is in (b), along with the 0%, 50% and 95% quantiles of the distribution (vertical dashed lines).

also had a non-negligible impact for this dataset (Table 2). Critically, local retention estimation was only possible for this data because the total size of the settler pool was experimentally assessed over the course of the study.

## Estimation of the number of POPs

The distribution of LOD values for putative POPs for real damselfish recruits from New Caledonia closely follows that of simulated *false* POPs (compare gray bars and dashed red curve in Figure 7a), indicating that the vast majority of the 200 recruits examined do not come from a genotyped parent. The commonly used method of identifying POPs based on a LOD threshold that balances type I and type II error rates finds that there is one potential POP with a LOD above the threshold. However, the type I error rate is 0.78%, corresponding to 1.56 false POPs in a sample of 200 recruits, so one would conclude that the likelihood of having observed a true POP is small.

The Bayesian approach to parentage analysis described in Equations (6)–(9) confirms these results, but also allows one to estimate a distribution for self-recruitment directly from the LODs and quantify the likelihood of having observed a given number of POPs in the sample. The 95% CI for self-recruitment for this sample is (0% 2.7%) (Fig. 7b). This is close to, but significantly lower than, the CI of (7% 20%) estimated for the same population using TRAIL. This difference is likely attributable to bias in the non-random sample of 200 recruits and/or differences in the sampling protocol between TRAIL and genetics (e.g. genotyping once per year versus TRAIL marking each month of the spawning season). Randomly drawing values for φ from the distribution in Figure 7b, calculating for each φ the probability each larval fish in the sample is a true POP using Equation (9) and then ensemble averaging the combined probability that there are no true POPs in the sample over all values of φ yields that there is $<1\%$ probability of having a non-zero number of POPs in the larval sample, consistent with results based on a LOD threshold plus the type I error rate.

## Discussion

The methods developed here, using a combination of mark-recapture/site-occupancy Bayesian models, and bootstrapping over potential values of independently-estimated model input parameters, demonstrate the integration of multiple sources of uncertainty in marine larval connectivity studies. We use these methods to estimate uncertainty in two important connectivity indices: self-recruitment and local retention. Statistical frameworks for future developments in the field are also described, such as using multinomial models to assess connectivity in TRAIL studies with multiple micro-chemical markers or, in parentage-analysis studies, using the relative observation rates of settlers with one versus two known parents to assess population reproductive structure or estimate the fraction of adults marked. The methods developed here are likely to become essential as larval dispersal studies are used increasingly in the context of marine management, which often requires examining species with larger population sizes and estimating connectivity for multiple sites and time periods. In these cases, empirical studies of marine connectivity will very likely involve non-exhaustive sampling of adults and settlers and, therefore, require accurate assessment of uncertainty. Furthermore, management decisions will ideally be based on comparisons across multiple connectivity estimation approaches (e.g. empirical and numerical), for which a comprehensive understanding of uncertainty is essential.

Surprisingly, uncertainty due to the small size of the sample of settlers collected was the dominant source of uncertainty for connectivity estimates for the majority of the published studies examined. For example, widths of 95% CI for relative connectivity (or self-recruitment) due to small sample size uncertainty exceeded 0.5 for many published connectivity results. As relative connectivity must be between 0 and 1, this represents a very significant amount of uncertainty, complicating using individual self-recruitment results to conclude that marine populations are relatively closed or open (however, consistency among a suite of estimates or studies may still allow one to draw general conclusions about openness in marine populations).

Small sample size uncertainty was dominant despite the considerable uncertainty in the fraction of eggs marked in many studies. Exhaustive marking of mature adults *and* collection of settlers can remove this uncertainty, but uncertainty rapidly approaches that of a theoretical infinite settler pool if either of these experimental steps is not exhaustive. Collecting a larger number of settlers also reduces small sample size uncertainty, and, therefore, high-throughput genotyping techniques capable of rapidly genotyping large numbers of individuals may help reduce this uncertainty in the future. Nevertheless, given the increasingly complex questions asked of empirical larval dispersal studies, small sample size is likely to remain an important source of measurement uncertainty, happily, relatively easy to assess.

Critically, many studies do not provide enough detail to properly assess uncertainty in relative connectivity estimates and/or permit calculating true elements of the connectivity matrix, such as local retention. In many cases, it is unclear whether or not spawners and settlers were exhaustively sampled. Target species reproductive behaviour and output were often not discussed despite their importance for estimating uncertainty in the fraction of eggs marked and assessing the need for correcting this fraction for dual mark transmission (in genetic studies). Total settlement at the destination site was rarely assessed, making it impossible to convert relative connectivity values into true elements of the connectivity matrix. Systematically including in marine larval connectivity studies indicators of exhaustiveness of settler collection and a review of target species reproductive behaviour (see next paragraph) is relatively easy and would considerably enhance the utility of study results. Assessing reproductive output and the absolute size of the settler pool often requires additional fieldwork, but will produce more robust and useful connectivity estimates.

Equal probability of mating between genotyped and non-genotyped individuals is an important underlying assumption of many parentage-based larval connectivity studies. Whether this assumption applies to a system depends on the structure of mating between individuals in the population and the sampling strategy. For example, colonial damselfish typically only mate with other individuals in the colony, so genotyping all individuals from a damselfish colony should only produce larvae with two known parents. In this case, only maternal mark transmission is important, and the fraction of "marked" eggs is simply equal to the fraction of genotyped adult females, as in TRAIL studies. On the other hand, dual maternal and paternal mark transmission may accurately represent aggregative spawners. Not knowing the extent to which either of these two limiting cases applies to a given study is potentially a large source of uncertainty in connectivity estimates (as much as a factor of 2 error in $p$). Therefore, understanding reproductive behaviour is essential to parentage-based connectivity studies, and basic information regarding reproductive behaviour should be included in future studies.

Overall, we hope that this study and the ConnMatTools R package implementing the statistical methods developed here will contribute to producing more robust and useful marine connectivity estimates in the future. We feel strongly that this detailed approach to uncertainty is needed if studies of marine larval connectivity are going to make an important contribution to our understanding of marine population dynamics and management of marine ecosystems.

## Supplementary data

Supplementary material is available at the *ICESJMS* online version of the article.

## Acknowledgements

We would like to acknowledge Glenn Almany, whose work inspired many aspects of this study. This article also benefited

from conversations with Len Thomas and Mark R. Christie. We also thank the handling editor and three anonymous reviewers for their constructive comments that improved the article.

## Funding

Funding was provided by the COMPO project (Connectivity Of Marine POpulations, www.compo.ird.fr) through a grant from the French National Research Agency (ANR) no. 2010 JCJC 1701 01. This is contribution number 3585 of the Virginia Institute of Marine Science, College of William and Mary.

## References

Almany, G. R., Berumen, M. L., Thorrold, S. R., Planes, S., and Jones, G. P. 2007. Local replenishment of coral reef fish populations in a marine reserve. Science, 316: 742–744.

Berger, J. O., Bernardo, J. M., and Sun, D. 2015. Overall objective priors. Bayesian Analysis, 10: 189–221.

Berumen, M. L., Almany, G. R., Planes, S., Jones, G. P., Saenz-Agudelo, P., and Thorrold, S. R. 2012. Persistence of self-recruitment and patterns of larval connectivity in a marine protected area network. Ecology and Evolution, 2: 444–452.

Botsford, L., Brumbaugh, D., Grimes, C., Kellner, J., Largier, J., O'Farrell, M., Ralston, S., et al. 2009. Connectivity, sustainability, and yield: bridging the gap between conventional fisheries management and marine protected areas. Reviews in Fish Biology and Fisheries, 19: 69–95.

Botsford, L. W., Micheli, F., and Hastings, A. 2003. Principles for the design of marine reserves. Ecological Applications, 13: S25–S31.

Burgess, S. C., Nickols, K. J., Griesemer, C. D., Barnett, L. A. K., Dedrick, A. G., Satterthwaite, E. V., Yamane, L., et al. 2014. Beyond connectivity: how empirical methods can quantify population persistence to improve marine protected-area design. Ecological Applications, 24: 257–270.

Christie, M. R., Tennessen, J. A., and Blouin, M. S. 2013. Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. Bioinformatics, 29: 725–32.

Christie, M. R., Tissot, B. N., Albins, M. A., Beets, J. P., Jia, Y., Ortiz, D. M., Thompson, S. E., et al. 2010. Larval Connectivity in an Effective Network of Marine Protected Areas. PLoS One, 5: e15715.

Cowen, R. K., Paris, C. B., and Srinivasan, A. 2006. Scaling of connectivity in marine populations. Science, 311: 522–527.

Cuif, M., Kaplan, D. M., Fauvelot, C., Lett, C., and Vigliola, L. 2015. Monthly variability of self-recruitment for a coral reef damselfish. Coral Reefs, 34: 759–770.

D'Aloia, C. C., Bogdanowicz, S. M., Francis, R. K., Majoris, J. E., Harrison, R. G., and Buston, P. M. 2015. Patterns, causes, and consequences of marine larval dispersal. Proceedings of the National Academy of Sciences of the United States of America, 112: 13940–13945.

D'Aloia, C. C., Bogdanowicz, S. M., Majoris, J. E., Harrison, R. G., and Buston, P. M. 2013. Self-recruitment in a Caribbean reef fish: a method for approximating dispersal kernels accounting for seascape. Molecular Ecology, 22: 2563–2572.

Garavelli, L., Kaplan, D. M., Colas, F., Stotz, W., Yannicelli, B., and Lett, C. 2014. Identifying appropriate spatial scales for marine conservation and management using a larval dispersal model: The case of Concholepas concholepas (loco) in Chile. Progress in Oceanography, 124: 42–53.

Gerber, S., Chabrier, P., and Kremer, A. 2003. FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Molecular Ecology Notes, 3: 479–481.

Gunderson, D. R. 1997. Trade-off between reproductive effort and adult survival in oviparous and viviparous fishes. Canadian Journal of Fisheries and Aquatic Sciences, 54: 990–998.

Hamilton, S. L., Regetz, J., and Warner, R. R. 2008. Postsettlement survival linked to larval life in a marine fish. Proceedings of the National Academy of Sciences of the United States of America, 105: 1561–1566.

Harrison, H. B., Williamson, D. H., Evans, R. D., Almany, G. R., Thorrold, S. R., Russ, G. R., Feldheim, K. A., et al. 2012. Larval Export from Marine Reserves and the Recruitment Benefit for Fish and Fisheries. Current Biology, 22: 1023–1028.

Hess, M. A., Rabe, C. D., Vogel, J. L., Stephenson, J. J., Nelson, D. D., and Narum, S. R. 2012. Supportive breeding boosts natural population abundance with minimal negative impacts on fitness of a wild population of Chinook salmon. Molecular Ecology, 21: 5236–5250.

Jones, A. G., Small, C. M., Paczolt, K. A., and Ratterman, N. L. 2010. A practical guide to methods of parentage analysis: TECHNICAL REVIEW. Molecular Ecology Resources, 10: 6–30.

Jones, G. P., Milicich, M. J., Emslie, M. J., and Lunow, C. 1999. Self-recruitment in a coral reef fish population. Nature, 402: 802–804.

Jones, G. P., Planes, S., and Thorrold, S. R. 2005. Coral Reef Fish Larvae Settle Close to Home. Current Biology, 15: 1314–1318.

Kaplan, D. M., Planes, S., Fauvelot, C., Brochier, T., Lett, C., Bodin, N., Le Loc'h, F., et al. 2010. New tools for the spatial management of living marine resources. Current Opinion in Environmental Sustainability, 2: 88–93.

Lett, C., Nguyen-Huu, T., Cuif, M., Saenz-Agudelo, P., and Kaplan, D. M. 2015. Linking local retention, self-recruitment, and persistence in marine metapopulations. Ecology, 96: 2236–2244.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., and Hines, J. E. 2005. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press, San Diego. 344 pp.

Nickols, K. J., White, J. W., Largier, J. L., and Gaylord, B. 2015. Marine population connectivity: reconciling large-scale dispersal and high self-retention. The American Naturalist, 185: 196–211.

Planes, S., Jones, G. P., and Thorrold, S. R. 2009. Larval dispersal connects fish populations in a network of marine protected areas. Proceedings of the National Academy of Sciences of the United States of America, 106: 5693–5697.

Saenz-Agudelo, P., Jones, G. P., Thorrold, S. R., and Planes, S. 2011. Connectivity dominates larval replenishment in a coastal reef fish metapopulation. Proceedings of the Royal Society B: Biological Sciences, 278: 2954–2961.

Sala, E., Aburto-Oropeza, O., Paredes, G., Parra, I., Barrera, J. C., and Dayton, P. K. 2002. A general model for designing networks of marine reserves. Science, 298: 1991–1993.

Schunter, C., Pascual, M., Garza, J. C., Raventos, N., and Macpherson, E. 2014. Kinship analyses identify fish dispersal events on a temperate coastline. Proceedings of the Royal Society of London B: Biological Sciences, 281: 20140556.

Schwarz, C. J. 2001. The Jolly-Seber model: more than just abundance. Journal of Agricultural, Biological, and Environmental Statistics, 6: 195–205.

Strathmann, R. R. 1990. Why life histories evolve differently in the sea. American Zoologist, 30: 197–207.

Thorrold, S. R., Jones, G. P., Planes, S., and Hare, J. A. 2006. Transgenerational marking of embryonic otoliths in marine fishes using barium stable isotopes. Canadian Journal of Fisheries and Aquatic Sciences, 63: 1193–1197.

*Handling editor: Manuel Hidalgo*