Estimating dispersal kernels using genetic parentage data

Authors: Michael Bode [1,2] *, Nick Outram [1], Geoffrey P. Jones [2,3]

Affiliations

1. ARC Centre of Excellence for Environmental Decisions, School of Botany, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia.
2. Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville 4811 QLD Australia.
3. College of Marine and Environmental Sciences, James Cook University, Townsville 4811 QLD Australia.

* Corresponding author

Email: mbode@unimelb.edu.au

Phone: +61 4 14 108 439

Abstract

Dispersal kernels are a standard method for describing and predicting the relationship between dispersal strength and distance. Kernels allow observations of a limited number of dispersal events to be extrapolated across a wider landscape, and form the basis of a wide range of theories and methods in ecology and conservation. Genetic parentage data are an increasingly common source of dispersal information, particularly for species where dispersal is difficult to observe directly. It is now routinely applied to coral reef fish, where larvae disperse over many kilometers and are too small to follow directly. However, it is not straight forward to estimate dispersal kernels from parentage data, and while a number of different methods have been published, these have been examined systematically, and each has substantial limitations. Here we develop and test a new statistical estimator for fitting dispersal kernels to parentage data. The method incorporates a series of factors omitted in previous methods: the partial sampling of adults and juveniles on sampled reefs; accounting for dispersers from unsampled reefs; and post-settlement processes (e.g., density dependent mortality) that follow dispersal but precede parentage sampling. Power analyses indicate that the highest levels of sampling currently used for reef fishes is sufficient to fit accurate dispersal kernels. Sampling is best distributed equally between adults and juveniles, and over an area of less than twice the mean dispersal distance (particularly when sampling is limited). Unlike previous methods, accounting for unsampled adults – both on partially-sampled and unsampled patches – is essential for a precise and unbiased estimate of dispersal.

Keywords: Connectivity; parentage analysis; self-recruitment; ghost patches.

Introduction

The pattern and strength of dispersal play a defining role in ecology and evolution, particularly in ecosystems where suitable habitat is patchily distributed (Clobert *et al.* 2001; Bullock *et al.* 2002). However, dispersal is a difficult process to observe, particularly when dispersing individuals are small, numerous and hard to follow (Cowen & Sponaugle 2009; Jones *et al.* 2009). A wide variety of approaches have been developed to measure dispersal events. These include tracking individual trajectories using radio or GPS devices; mark-recapture across multiple patches; and inverse modelling using the location and species identity of propagules and adults (Clobert *et al.* 2001; Bullock *et al.* 2002; Clark *et al.* 2008; Broquet & Petit 2009). In recent years, population and individual-based genetic methods have begun to offer unique and powerful insights into dispersal (Jones *et al.* 2009; Broquet & Petit 2009).

Hypervariable molecular markers allow sampled juveniles to be assigned to sampled parents, and can thereby provide individual dispersal vectors. The approach was first used to estimate dispersal and gene flow in plant communities (Ellstrand & Marshall 1985), has been applied to birds (Woltmann *et al.* 2012), and mammals (Telfer *et al.* 2003), and is increasingly being used to observe larval dispersal events in marine fish metapopulations, particularly on coral reefs (Jones *et al.* 2005b; Almany *et al.* 2007; Planes *et al.* 2009; Buston *et al.* 2012; Harrison *et al.* 2012; Almany *et al.* 2013; D'Aloia *et al.* 2015). Parentage analysis has been used to conclusively answer open ecological questions (Almany *et al.* 2007; Jones *et al.* 2009), and provide suggestive evidence about the role of connectivity in spatial management (Harrison *et al.* 2012). However, parentage assignment data often report a small number of dispersal events, at a particular time and in a subset of locations. Many questions in ecology, and many decision-support tools in conservation, require a more expansive description of dispersal.

We therefore require statistically robust methods for extrapolating limited parentage data to predictions about dispersal in the broader landscape. The most common approach at present is to estimate how dispersal strength relates to the distance between habitat patches, since isolation-by-distance is a natural way to conceive of dispersal dynamics. Operationally, this involves fitting a parametric relationship – a dispersal kernel – to a set of observed dispersal events (Clobert *et al.* 2001; Nathan *et al.* 2012). Previous analyses have fit dispersal kernels to parentage datasets using a range of different methods (Jones *et al.* 2005a; Jones & Muller Landau 2008; Buston *et al.* 2012; Almany *et al.* 2013; Hopf *et al.* 2015; D'Aloia *et al.* 2015). At their most basic, these methods involve calculating the average distance travelled by juveniles that were sampled and assigned to parents (Woltmann *et al.* 2012), or fitting a regression line (Telfer *et al.* 2003; Buston *et al.* 2012) directly to the distances travelled. In essence, these methods use each parentage assignment as a datapoint relating distance to dispersal strength. While this assumption is broadly reasonable, the raw number of assignments cannot be simply regressed against inter-patch distance. Parentage data are the result of dispersal, but they also reflect other ecological processes (e.g., post-recruitment mortality) and the distribution and

intensity of the sampling process that generated them. Parentage data can only inform dispersal once these factors are corrected for.

All methods currently used to fit dispersal kernels are subject to limitations. First, Second, kernels predict the proportion of dispersing individuals that leave one patch and travel to another (dispersers, such as seeds or larvae). In contrast, parentage analyses measure the number of juveniles on one patch who arrived from another (recruits, such as seedlings or juveniles). Because this latter measurement is made after some level of post-settlement mortality has occurred, dispersal cannot be directly calculated by a simple fit to recruits (Fig 1B).

The fit of dispersal kernels to parentage data will theoretically be influenced by a set of factors that are generally not included in current fitting methods. (1) Even the most extensive parentage datasets contain only subsamples of the adult populations. The data therefore contains a large numbers of juveniles – generally the majority – that cannot be assigned to any parents. These unassigned juveniles could be the offspring of adults on "ghost populations", that is, patches that are completely unsampled (Beerli 2004; Wang 2014). They could also be the offspring of unsampled adults from patches where adults were only partially sampled (Jones & Ardren 2003). Both options must be considered by the method. (2) Parentage datasets contain finite samples of the juvenile populations, and the datasets vary spatially in sampling intensity (Jones & Muller Landau 2008). (3) These juvenile samples were not collected immediately following the dispersal phase, and therefore are the result of dispersal, filtered by post-settlement mortality processes (Moran & Clark 2012). Including any of these factors can dramatically change the interpretation of a given parentage sample (Figure 1A).

All previous estimates of larval dispersal kernels for coral reef fish have ignored at least one of these factors, making their estimates either biased or over-confident. In this paper, we propose a novel likelihood estimator for dispersal kernels that incorporates each of these factors, and demonstrate its application to a case study of reef fish larval dispersal on the Great Barrier Reef. We use simulated datasets to investigate the power and statistical properties of our estimator under logistical constraints, which limit the number of juveniles sampled, the proportion of adults sampled at each patch, and the total number of patches sampled.

A likelihood function for parentage-based dispersal data

The metapopulation comprises P habitat patches, each with a population of N_i individuals (we assume an approximately equal sex ratio on all patches). The metapopulation consists of individuals in three stages. "Adults" are reproductively mature individuals that produce dispersers (e.g., larvae) that move according to a dispersal kernel. If these offspring arrive at a habitat patch, they become "settlers", and attempt to recruit to the local population. At this point they suffer density-dependent mortality; those that survive become "juveniles" and may be sampled for parentage.

Adult sampling takes place on a subset $\mathbf{S_A}$ of these patches, where $\mathbf{S_A}$ has s_A elements. Juveniles are sampled from a subset $\mathbf{S_J}$, with s_J elements ($\mathbf{S_A}$ and $\mathbf{S_J}$ are often the same set). A proportion π_i of the adults on reef i are sampled, with $\pi_i = 0$ for ghost patches. Each sampled juvenile is either assigned to a sampled adult, or classified as having unknown parentage. This count data populates matrix **M**, with the columns indicating the patch where the juvenile was sampled, and the row indicating the natal patch. The final row contains the number of unassigned juveniles, which can dominate the matrix. The count data in **M** are a sample from the juveniles on each patch, which are themselves samples of the settlers. We assume that both the juveniles and settlement pools are large, so that we can model recruitment and juvenile collection as samples with replacement. We also assume that each disperser, regardless of its origin, has an equal probability of successfully settling (e.g., there is no local adaptive advantage (Warner 1997) or cost to long-duration dispersal (Burgess *et al.* 2013a)).

The settlement pool on each patch is determined by the dispersal kernel $\rho(k, d_{ij})$, which assumes dispersal is isotropic, spatially invariant, and based on the distance d_{ij} between patches i and j. The kernel is a probability density function whose shape is defined by the parameter set k. One commonly-used kernel is the generalised Gaussian function (Largier 2003; Bode *et al.* 2011; Nathan *et al.* 2012):

$$\rho(k, d) = \exp\left[-(kd)^{\theta}\right]$$

Eq. 1

where θ is a shape parameter that yields the standard Gaussian when $\theta = 2$, and the Laplacian or negative exponential when $\theta = 1$. Normally these kernels have standardizing coefficients to ensure that they integrate to one. However, because we were interested in the proportional make-up of the settlement pool, any such coefficients will occur in both the numerator and denominator (see Eq. 2).

According to a particular dispersal kernel, the proportion of the settlers on patch j that come from sampled adults on reef i will be:

$$q_{ij}(k) = \frac{(1 - (1 - \pi_i)^2)\, N_i\, \rho(k, d_{ij})}{\sum_{m=1}^{P} N_m\, \rho(k, d_{mj})}.$$

Eq. 2

The numerator in Eq. 2 calculates the number of settlers who were created on patch i by at least one sampled parent, who dispersed to patch j. The denominator divides this by the total number of settlers arriving at patch j, from both sampled and unsampled adults, turning it into a probability. This division recreates the effect of any density-dependent settlement mortality process that is neutral to the source of the settlers. Both the numerator and denominator could be modified by the per-capita fecundity of the females, the proportion of successfully fertilised eggs, and the mortality during the dispersal phase. However, if we assume that these do not vary between patches, they do not alter the fit. Equation 2 depends heavily on the population size on potential source patches, N_j (Fig 1a). These

must therefore be either sampled, or estimated using observed densities on comparable sampled habitat.

In most parentage datasets, a large proportion of the juveniles cannot be assigned to any of the sampled adults. The dispersal kernel also predicts this proportion, which either come from ghost patches, or from unsampled adults on sampled patches:

$$\overline{q}_j(k) = 1 - \sum_{i=1}^{P} q_{ij}(k) \ = \frac{\sum_{i=1}^{P}(1-\pi_i)^2\, N_i\, \rho\big(k, d_{ij}\big)}{\sum_{m=1}^{P} N_m\, \rho\big(k, d_{mj}\big)}.$$

Eq. 3

Applying Eq. 2 and Eq. 3 to the parentage matrix **M**, the log likelihood of observing the set of attributed and unattributed samples, given dispersal kernel $\rho(k, d)$ is therefore:

$$LL(k|\mathbf{M}) = \sum_{j=1}^{P} \ln\binom{R_{ij}}{\mathbf{M}_{ij}} + \sum_{j\in\mathbf{S}_J}\left[\mathbf{M}_{Xj}\ln\left(\overline{q}_j(k)\right) + \sum_{i=1}^{P}\mathbf{M}_{ij}\ln\left(q_{ij}(k)\right)\right],$$

Eq. 4

Where $R_{ij} = \sum_i \mathbf{M}_{ij}$, the total number of recruits sampled on patch *j*. The index variable X in $\mathbf{M}_{Xj}$ refers to the final row in the parentage matrix (i.e., unallocated juveniles, so $X = s_A + 1$). The very first summation – a sum over multinomial coefficients – does not change for different parameter sets, and should therefore be ignored since it can be large enough to pose numerical issues. Confidence intervals can be generated by repeatedly finding the maximum likelihood fit for bootstrap resamples of the data, using source reefs as resampling units.

Power analyses for a reef fish case-study

The collection and analysis of parentage data is both expensive and multi-dimensional: a patch must be visited, and adults and/or juveniles must be caught, biopsied (sometimes destructively), genotyped and then genetically matched. Total sampling effort must be shared between adults and juveniles, and across multiple patches. We perform a series of simulation-based power analyses to assess how these decisions affect the accuracy of the best-fit dispersal kernel. To do this, we (1) simulate populations and their parentage relationships, (2) create subsamples of juveniles and adults, and (3) analyse them using the fitting method described above.

<u>Simulating the populations:</u> We use a spatially explicit metapopulation model of reef fish on reefs in the Cairns Management Region of the Great Barrier Reef to generate population sizes and parentage relationships for sampling (Figure 1; full details on the population model can be found in Bode *et al.* 2012). Populations on each reef in the metapopulation are governed by age-structured dynamics, with density-independent survival rates for adults, and density-dependent mortality of settling larvae

based on a Beverton-Holt function. The species' life-history is broadly parameterised for coral trout (genus *Plectopomus*) species, a target of recreational fishers and a focus of previous genetic parentage analysis and larval kernel estimation (Harrison *et al.* 2012; Almany *et al.* 2013; Hopf *et al.* 2015). Reproduction is followed by an obligate larval dispersal phase, which we model using a particular dispersal kernel. By keeping track of both parents and juveniles, we simulate a population dataset with parentage relations that can be repeatedly sampled. Although only a subset of reefs are sampled, we generate a single generation of parentage data for the whole metapopulation of 320 reefs over 500 km of coastline, since we are interested in the effects of including or ignoring unsampled populations and unattributed juveniles. A large proportion of our juveniles are therefore unassigned.

Creating & fitting parentage data: Using this simulated data, we sample a fixed number of individuals (adults and juveniles put together) from a subset of the reefs within a contiguous region (Figure 1). From each sampled reef, we sample a given number of individuals from the post-density-dependent juvenile population, and a proportion of the adult population. Adult coral trout on the GBR exist at densities of around 3,500 km^{-2} (Cornish & Kiwi 2006), and we use this density to translate a proportional adult sample into a number of adult individuals (e.g., 1% of adults on a 1 km^2 reef is equivalent to 35 individuals).

We take each juvenile in turn and identify its source reef, if either of its parent were sampled. Although individual larvae will depart from and arrive at specific locations within each reef, we measure dispersal distances using the centroids of the reef perimeters, since the inter-reef distance is generally much larger than the size of individual reefs. While incorporating the precise locations of the adults and juveniles may hypothetically offer a more precise parameterisation, dispersers are not necessarily spawned at the location where an adult was sampled (e.g., *Plectopomus* species often spawn in aggregations), nor did juveniles necessarily settle at the precise location where they were sampled (White 2015). We aggregate this data to create a simulated parentage matrix.

We use our estimator (Eq. 4) to find the maximum likelihood parameter fit to these data, using the correct kernel functional form. We repeat the sampling and fitting procedure 1,000 times to estimate confidence intervals, taking each sample from a different part of the metapopulation to average over the effects of a heterogeneous patch distribution (Figure 1). We then calculate and report the mean dispersal distance of the kernel, m, which is a more comprehensible and widely reported quantity than the kernel parameter.

Power analyses: Our first power analysis assesses the performance of parentage analyses undertaken at the intensity and scale of contemporary research efforts. The case-studies in Table 2 involve sampling of between 1,000 and 5,000 individuals (adults and juveniles) across 19 to 66 reefs. We calculate their ability to recreate the mean dispersal distance of two hypothetical species, with exponential dispersal kernels $\rho(k, d) = \exp[-kd]$, where $k = 0.08$ and $k = 0.36$ ($m = 12.5$ km and $m = 2.78$ km respectively). We apply three scenarios of sampling intensity – low sampling effort

(1,000 individuals across 20 patches); intermediate effort (2,000 individuals across 40 patches); and high effort (7,000 individuals across 70 patches).

For parentage analyses, sampling effort is distributed along three primary dimensions: (1) the number of sampled juveniles, (2) the number of sampled adults, and (3) the number of sampled patch populations (the distribution of sampling effort across space). Total sampling effort is the product of these three dimensions; different distributions will result in different estimation accuracy, and the best distribution may vary with the dispersal kernel. We outline four more power analyses below, and on the basis of the results make a series of recommendations about how a given sampling budget should be distributed to best estimate the dispersal kernel shape and the mean dispersal distance.

Our second power analysis considers how to best distribute sampling effort between juveniles and adults. A focus on sampling juveniles generates a larger number of dispersal events, however, more adult samples give a higher probability that each event will be usefully attributed to a source and destination patch. Assuming that both adults and juveniles are sampled from the same set of 25 patches, we sample the genetics of 2,500 individuals: 100 from each patch. We share this sampling effort between adults and juvenile individuals in different proportions, ranging from 10% adults (and therefore 90% juveniles) through to 90% adults. We repeat this analysis for 1,000 and 5,000 individual samples, and for two mean dispersal distances that roughly correspond to the examples in Table 2.

In our third power analysis, we consider the distribution of sampling effort across space. Sampling across a larger number of patches increases our ability to sample the kernel tail (D'Aloia *et al.* 2015), but it means that we sample fewer individuals on each patch, and therefore see more unassigned individuals. Assuming that sampling will be equally focused (50:50) on adults and juveniles, we sample the genetics of 2,500 individuals collected from between 5 and 50 different patches. Once again, we repeat this analysis for 1,000 and 5,000 individual samples, and for two mean dispersal distances. We note that, while the spatial scale of the sampling will scale on average with the number of patches (i.e., a larger number of patches is generally distributed over a larger area), the absolute scale of the sampling in kilometers will depend on the density of patches in the sample area. At the same time as we assess the distribution of sampling effort across space, we also calculate the performance of directly fitting a kernel to assigned juveniles (Telfer *et al.* 2003; Buston *et al.* 2012), ignoring all the factors we have discussed to this point. We simply calculate the maximum likelihood kernel fit, using the dispersal kernel to calculate the probability of each event (an assigned juvenile).

For our fourth analysis we consider the implications of incomplete sampling of adult populations on sampled patches. When a sampled juvenile cannot be assigned to one or both parents, the kernel fitting procedure must decide if it came from a ghost population (implying a long-distance disperser), or from unsampled adults on a sampled patch (implying a short-distance disperser). Low intensity adult sampling sends conflicting messages to the estimator because it makes both options a possibility.

We sample 500 juveniles across 25 reefs, and apply adult sampling proportions that range between 1% and 100%. We repeat these analyses for 250 and 1,000 juvenile samples, and for two dispersal distances. Note that, unlike the previous power analyses, we do not vary any other parameters (e.g., number of juveniles or patches sampled) as we increase the proportion of adults sampled. We therefore expect the fit to improve as the proportion increases, and are interested in the rate of this increase.

For our firth and final power analysis, we calculate the impact of ignoring ghost populations and unassigned juveniles into our fitting methods. We create parentage data for a set of 2,500 individuals, distributed equally across adults and juveniles on 25 reefs. We contrast the mean dispersal distance when unassigned juveniles are either included, as we propose in Eq. 4, or are simply ignored by fitting a dispersal kernel to the assigned samples only. We repeat this analysis with 1,000, 3500 and 5,000 individuals, and two mean dispersal distances.

Results

In our first power analysis, we calculate the expected performance of current levels of sampling. For a species with a mean dispersal distance of 12.5 km, low sampling effort (1,000 individuals across 20 patches) provides an estimate between 6.7 km and 84.6 km (53% to 676% of the true value); intermediate effort (2,000 individuals across 40 patches) provides an estimate between 8.4 km and 26 km (67% to 207%); and high effort (7,000 individuals across 70 patches) provides an estimate between 10.8 km and 14.4 km (86% to 115%). The results are more precise for the species with shorter mean dispersal distance of 2.78 km, but are still only reasonable for the highest intensity sampling. For this species, low sampling effort provides an estimate between 0 km and 3.6 km (0% to 130%); intermediate effort provides an estimate between 2.2 km and 3.3 km (80% to 119%); and high effort provides an estimate between 2.6 km and 2.9 km (94% to 106%). These results indicate that kernels are harder to estimate for species with longer dispersal distances, a conclusion that holds across all of our power analyses. To synopsise: All levels of sampling provide unbiased estimates of the mean dispersal distance, but only the highest intensity sampling provides reasonably narrow confidence intervals.

Our second power analysis considers the distribution of sampling effort between juveniles and adults. Overall, the performance of the estimator does not vary with the proportion of sampling effort allocated to juveniles and adults (Figure 3; *Supplementary Figure SX*). The best results are achieved by allocating the sampling effort evenly, but the outcomes are essentially the same, as long as allocations do not fall below 25% of sampling effort for either adults or juveniles. Datasets outside this range tend to underestimate the mean dispersal distance, particularly for longer distance dispersers.

The third analysis considers the distribution of sampling effort across space. The most accurate and precise estimates of dispersal distances are achieved by focusing sampling effort on a small number of

well-sampled patches. Sampling individuals over a wider area results in a less precise estimate of the kernel, for both long- and short-distance dispersers (Figure 4; *Supplementary Figure SX*). This is particularly true when the total sampling effort is small. Spreading a small number of samples over a large space results in enormous overestimates of the mean dispersal distance. For larger sampling budgets, the size of the sampling region has a smaller influence on the fit precision.

The fourth analysis addresses incomplete adult sampling. The results show that the estimator is unbiased, regardless of the proportion of the adult population sampled (Figure 5; *Supplementary Figure SX*), and is consistently precise if the proportion of the adult population sampled remains above 5%. However, once the proportion fell below this level, the performance of the fit declined dramatically, particularly for lower total sampling budgets. Unlike the previous figures, this power analysis keeps all other sampling decisions constant as it increases the proportion of adults sampled. These results would therefore not justify increasing the sampled adult proportion above 10%, since the additional adults would have little impact on the fit, but would reduce the number of juveniles or patches sampled.

The final power analysis measures the impact of ignoring ghost populations and unassigned juveniles when fitting kernels. Not including these populations has a negative effect on estimator performance, particularly when fewer individuals were sampled, and when the species had a short dispersal distance (Figure 6; *Supplementary Figure SX*). Specifically, ignoring unassigned juveniles creates a consistent underestimate of dispersal distance, most dramatically for shorter dispersers. The resulting fits were often very poor. In conditions when the inclusion of unsampled patches produced accurate and precise estimates of the dispersal kernel (i.e., 5,000 individuals sampled across 25 reefs), ignoring the effects of unsampled adults meant that the 95% confidence intervals no longer enclosed the true value.

Discussion

Our power analyses demonstrate that, when fitting dispersal kernels and estimating dispersal distance, it is essential to account for the presence of ghost populations and unsampled adults. Fitted relationships between distance and dispersal strength that do not consider these adults, and particularly direct fits of kernels to assigned juveniles, may be highly inaccurate. In addition, our results indicate that to accurately characterise a dispersal kernel using parentage analyses, researchers should ideally collect at least 5,000 samples from between 5 and 10 patches. The samples should be allocated as evenly as possible between adults and juveniles, while ensuring that at least 10% of the adults on each patch are sampled.

On the basis of these results, only some of the currently available parentage datasets are powerful enough to provide reasonably accurate estimates of species' dispersal kernels. While some studies have sampled in excess of 10% of the adults (e.g. Harrison et al 2012, Almany et al 2013), for others this is a much lower proportion or unknown (e.g.). Many have achieved sample sizes of xxxxx or

more (refs), but others have been based in very limited sampling (refs). For those papers explicitly calculating dispersal kernels, failure to take account of unsampled populations (e.g. ref), % adults sampled (e.g. refs) may have seriously impacted on dispersal estimates.

Some of our conclusions are qualitatively unsurprising. For example, greater total sampling effort produces better quality fits, as does sampling larger proportions of the adult population on each reef. Other conclusions were less immediately obvious. For example, we show that total effort is best distributed equally between adults and juveniles. However, a greater emphasis on adults would have returned fewer unassigned juveniles, which could plausibly have given better estimates of the kernel. Similarly, sampling is best concentrated on a relatively small number of sites, greatly increasing the likelihood of positive assignments. However, it was also intuitively possible that an intermediate number of patches would have performed best, since this would have given a broader cross-section of the kernel shape, including the tail.

While our results show that current datasets are large enough to accurately characterise dispersal kernels, this is only true if managers apply appropriate fitting methods. In particular, we focus on two improvements to previous approaches: first, we incorporate unassigned juveniles, both by including the influence of ghost patches, and by correcting for unsampled adults on partially sampled reefs. Although these factors are understood to be important (Beerli 2004; Jones *et al.* 2005a; Wang 2014), they have not yet been incorporated into kernel estimators for species in coral reef ecosystems, where both are ubiquitous. Second, we acknowledge that the juveniles in parentage datasets are sampled after post-settlement density dependence has occurred. As a result, the proportion of dispersers who travel from patch *i* to patch *j* (dispersal: the predictions of a kernel) is very different the proportion of juveniles on patch *j* who can be assigned to adults on patch *i* (recruitment: the observations of a parentage analysis). This difference between dispersal and recruitment was highlighted by Burgess *et al.* (2014), who questioned the utility of recruitment estimates for assessing population persistence or making management decisions. Our results demonstrate that a dispersal kernel estimator can translate data on recruitment (including self-recruitment) into estimates of proportional dispersal, and proportional local retention.

Although our method includes more factors than previous efforts, it nevertheless makes simplifying assumptions about the processes that generate parentage data. First, we assume that parentage assignment is completely accurate, but the allocation error of parent-offspring relationships for reef fish species has been estimated at < 5% for false-positives, and < 1% for false-negatives (Harrison *et al.* 2012). While very low, these mis-assignments will lower the precision of our estimates. Moreover, the amount of sampling effort directed at the adult populations will affect these rates (Harrison *et al.* 2013b; Christie 2013; see: Harrison *et al.* 2013a) independent of the effects of sample size on the kernel estimation. These facts should be incorporated into the fit statistic. The likely result will be an increase in the uncertainty of the estimates but not their bias, and a greater emphasis on sampling parents than is current seen in our results.

Second, the estimator assumes binomial sampling without replacement, implying that the settlement pool is much larger than the number of settlers, and that the number of juvenile samples is a small proportion of the total number of juveniles. Larger proportions would require a different sampling model, and some estimate of the total pool size (which are not available for either settlement or juveniles).

Third, we assume a great deal of spatial homogeneity in key dispersal processes – mortality in the pelagic environment and upon settlement; adult density and fecundity; and the strength of density dependence. Such assumptions are probably incorrect in most ecosystems, although evidence is generally scarce. For reef fish species, each of these processes is thought to vary in space, and/or with the origin of the dispersers (Ruttenberg *et al.* 2005; Sponaugle & Grorud-Colvert 2006; Burgess *et al.* 2013b; Hixon & Webster), although the amount and patterns of variation are unknown for most species in most locations. Our assumptions of homogeneity therefore reflect data limitations, and probably transfer the underlying variation into more imprecise estimates.

Finally, our estimator only predicts proportional dispersal kernels, rather than the absolute number of spawned individuals who travel a given distance. The latter dispersal kernels cannot be constructed without additional information – specifically, estimates of settlement numbers before density-dependent mortality occurs – a very difficult quantity to measure (Almany & Webster 2006).

Each of these assumptions and simplifications could each be corrected with alterations to the estimator, if appropriate data were available. However, a much broader concern is the applicability and utility of kernel descriptions of dispersal in ecology. This is particularly true for coral reef fish, which are highly dispersive species in an advective and turbulent environment. Kernels use smooth probability distributions, which essentially assume that the dispersal process is homogeneous and temporally consistent. Propagules radiate out from each natal patch in a spatially-invariant and isotropic pattern; where advection is included, it is modelled as a consistent displacement of this smooth pattern (Almany *et al.* 2013; D'Aloia *et al.* 2015). The strength of dispersal to nearby patches is determined entirely by the distance between them, but in reality the distribution of habitat patches in the local neighbourhood will influence the amount of dispersal, as locomotive individuals choose between different settlement options (Gerlach *et al.* 2007), or choose to delay settlement in the hope of finding more suitable habitat. Importantly, biophysical models of larval dispersal, based on a complex understanding of oceanographic processes and forced and validated by extensive data, predict dispersal patterns that are highly variable in both space and time, at multiple scales (James *et al.* 2002; Cowen 2006; Bode *et al.* 2012). It is possible that biophysical dispersal – if averaged over a sufficiently long timespan – approaches a smooth kernel (Cowen & Sponaugle 2009), and that long-term management decisions can be based on such time-averaged descriptions of dispersal. However, it is currently unclear whether ecological dispersal in general, and coral reef fish dispersal in particular, can be accurately described using dispersal kernels. Crucially, however, the inaccuracy of kernel descriptions of dispersal does not automatically preclude them being useful for management. If

smooth kernels capture essential elements of dispersal dynamics – for example, the mean distance travelled – then decisions based on these imperfect descriptions may still satisfy management objectives if they respond primarily to mean dispersal distance (Runge *et al.* 2011; Moore & RUNGE 2012).

The contrast between the biophysical and kernel descriptions of dispersal therefore creates an important challenge for spatial ecologists and conservation biologists. Mechanistic, biophysical models of dispersal are almost certainly more accurate than the kernel-based descriptions, and explicitly include factors (e.g., spatial variation; temporal fluctuations) that are known to fundamentally alter ecological processes and conservation management (Chesson & Warner 1981; Amarasekare 2003; Berkeley *et al.* 2010). However, most ecological (Bode *et al.* 2011; Okubo & Levin 2013), evolutionary (Skellam 1951; Levin *et al.* 2003; Nathan 2006) and management theory (Hastings & Botsford 2003; Neubert & Parker 2004; Arim *et al.* 2006; White *et al.* 2008; Botsford *et al.* 2009), and almost all of our spatial planning tools (Moilanen & Wintle 2006; Lehtomaki *et al.* 2009; Carroll *et al.* 2010; Laitila & Moilanen 2013), are based on dispersal kernels, rather than connectivity matrices. These two descriptions of dispersal must be reconciled, or the simpler, kernel-based description must be refuted. This is especially true for the understanding and management of reef fish biodiversity, where dispersal plays such a pivotal role in demography (Cowen 2002; Bode *et al.* 2006), community dynamics (Salomon *et al.* 2010; Bode *et al.* 2011), and spatial management (Hastings & Botsford 2003; Almany *et al.* 2013; Green *et al.* 2014). Our development and testing of accurate dispersal kernel fitting methods is an essential step in this process.

References

Almany, G.R. & Webster, M.S. (2006). The predation gauntlet: early post-settlement mortality in reef fishes. *Coral Reefs*, **25**, 19–22.

Almany, G.R., Berumen, M.L., Thorrold, S.R., Planes, S. & Jones, G.P. (2007). Local Replenishment of Coral Reef Fish Populations in a Marine Reserve. *Science*, **316**, 742–744.

Almany, G.R., Hamilton, R.J., Bode, M., Matawai, M., Potuku, T., Saenz-Agudelo, P., Planes, S., Berumen, M.L., Rhodes, K.L., Thorrold, S.R., Russ, G.R. & Jones, G.P. (2013). Dispersal of Grouper Larvae Drives Local Resource Sharing in a Coral Reef Fishery. *Current Biology*, **23**, 626–630.

Amarasekare, P. (2003). Competitive coexistence in spatially structured environments: a synthesis. *Ecology Letters*, **6**, 1109–1122.

Arim, M., Abades, S.R., Neill, P.E., Lima, M. & Marquet, P.A. (2006). Spread dynamics of invasive species. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 374–378.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*.

Berkeley, H.A., Kendall, B.E., Mitarai, S. & Siegel, D. (2010). Turbulent dispersal promotes species coexistence. *Ecology Letters*, **13**, 360–371.

Bode, M., Armsworth, P.R., Fox, H.E. & Bode, L. (2012). Surrogates for reef fish connectivity when designing marine protected area networks. *Marine Ecology Progress Series*, **466**, 155–166.

Bode, M., Bode, L. & Armsworth, P.R. (2011). Different dispersal abilities allow reef fish to coexist. *Proceedings of the National Academy of Sciences*, **108**, 16317–16321.

Bode, M., Bode, L. & Armsworth, P.R. (2006). Larval dispersal reveals regional sources and sinks in the Great Barrier Reef. *Marine Ecology Progress Series*, **308**, 17–25.

Botsford, L.W., White, J.W., Coffroth, M.A., Paris, C.B., Planes, S., Shearer, T.L., Thorrold, S.R. & Jones, G.P. (2009). Connectivity and resilience of coral reef metapopulations in marine protected areas: matching empirical efforts to predictive needs. *Coral Reefs*, **28**, 327–337.

Broquet, T. & Petit, E.J. (2009). Molecular Estimation of Dispersal for Ecology and Population Genetics. *dx.doi.org.ezp.lib.unimelb.edu.au*, **40**, 193–216.

Bullock, J.M., Kenward, R.E. & Hails, R.S. (2002). *Dispersal ecology: 42nd symposium of the British ecological society*.

Burgess, S.C., Bode, M. & Marshall, D.J. (2013a). Costs of dispersal alter optimal offspring size in patchy habitats: combining theory and data for a marine invertebrate (C. Fox, Ed.). *Functional Ecology*, **27**, 757–765.

Burgess, S.C., Bode, M. & Marshall, D.J. (2013b). Costs of dispersal alter optimal offspring size in patchy habitats: combining theory and data for a marine invertebrate (C. Fox, Ed.). *Functional Ecology*, **27**, 757–765.

Burgess, S.C., Nickols, K.J., Griesemer, C.D., Barnett, L.A.K., Dedrick, A.G., Satterthwaite, E.V., Yamane, L., Morgan, S.G., White, J.W. & Botsford, L.W. (2014). Beyond connectivity: how empirical methods can quantify population persistence to improve marine protected-area design. *dx.doi.org.ezp.lib.unimelb.edu.au*, **24**, 257–270.

Buston, P.M., Jones, G.P., Planes, S. & Thorrold, S.R. (2012). Probability of successful larval dispersal declines fivefold over 1 km in a coral reef fish. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1883–1888.

Carroll, C., Dunk, J.R. & Moilanen, A. (2010). Optimizing resiliency of reserve networks to climate change: multispecies conservation planning in the Pacific Northwest, USA. *Global Change Biology*, **16**, 891–904.

Chesson, P.L. & Warner, R.R. (1981). Environmental variability promotes coexistence in lottery competitive systems. *The American Naturalist*, **117**, 923–943.

Christie, M.R. (2013). Bayesian parentage analysis reliably controls the number of false assignments in natural populations. *Molecular Ecology*, **22**, 5731–5737.

Clark, J.S., Silman, M., Kern, R., Macklin, E. & HilleRisLambers, J. (2008). SEED DISPERSAL NEAR AND FAR: PATTERNS ACROSS TEMPERATE AND TROPICAL FORESTS. *dx.doi.org.ezp.lib.unimelb.edu.au*, **80**, 1475–1494.

Clobert, J., Danchin, E., Dhondt, A.A. & Nichols, J.D. (2001). *Dispersal*.

Cornish, A. & Kiwi, L.K. (2006). *Plecropomus leopardus. Species Report for the IUCN Red List of Threatened Species*. IUCN, Gland, Switzerland.

Cowen, R.K. (2002). Larval dispersal and retention and consequences for population connectivity. *Coral reef fishes: dynamics and diversity in a complex ecosystem* (ed P. Sale), pp. 149–170. Academic

Press, San Diego.

Cowen, R.K. (2006). Scaling of Connectivity in Marine Populations. *Science*, **311**, 522–527.

Cowen, R.K. & Sponaugle, S. (2009). Larval Dispersal and Marine Population Connectivity. *Annual Review of Marine Science*, **1**, 443–466.

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

Ellstrand, N.C. & Marshall, D.L. (1985). Interpopulation gene flow by pollen in wild radish, Raphanus sativus. *American Naturalist*.

Gerlach, G., Atema, J., Kingsford, M.J., Black, K.P. & Miller-Sims, V. (2007). Smelling home can prevent dispersal of reef fish larvae. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 858–863.

Green, A.L., Fernandes, L. & Almany, G. (2014). Designing marine reserves for fisheries management, biodiversity conservation, and climate change adaptation. *Coastal ….*

Harrison, H.B., Saenz-Agudelo, P., Planes, S., Jones, G.P. & Berumen, M.L. (2013a). On minimizing assignment errors and the trade-off between false positives and negatives in parentage analysis. *Molecular Ecology*, **22**, 5738–5742.

Harrison, H.B., Saenz-Agudelo, P., Planes, S., Jones, G.P. & Berumen, M.L. (2013b). Relative accuracy of three common methods of parentage analysis in natural populations. *Molecular Ecology*, **22**, 1158–1170.

Harrison, H.B., Williamson, D.H., Evans, R.D., Almany, G.R., Thorrold, S.R., Russ, G.R., Feldheim, K.A., van Herwerden, L., Planes, S., Srinivasan, M., Berumen, M.L. & Jones, G.P. (2012). Larval Export from Marine Reserves and the Recruitment Benefit for Fish and Fisheries. *Current Biology*, **22**, 1023–1028.

Hastings, A. & Botsford, L.W. (2003). Comparing designs of marine reserves for fisheries and for biodiversity. *Ecological Applications*, **13**, S65–S70.

Hixon, M. & Webster, M.S. Density dependence in reef fish populations. *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed P.F. Sale), pp. 303–325. Academic Press, San Diego, California.

Hopf, J.K., Jones, G.P. & Willamson, D.H. (2015). Fishery consequences of marine reserves: short-term pain for longer-term gain. *Ecological Applications*.

James, M.K., Armsworth, P.R., Mason, L.B. & Bode, L. (2002). The structure of reef fish metapopulations: modelling larval dispersal and retention patterns. *Proceedings of the Royal Society B: Biological Sciences*, **269**, 2079–2086.

Jones, A.G. & Ardren, W.R. (2003). Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Jones, F.A. & Muller Landau, H.C. (2008). Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. *Journal of Ecology*, **96**, 642–652.

Jones, F.A., Chen, J., Weng, G.J. & Hubbell, S.P. (2005a). A Genetic Evaluation of Seed Dispersal in the Neotropical Tree Jacaranda copaia(Bignoniaceae). *The American Naturalist*, **166**, 543–555.

Jones, G.P., Almany, G.R., Russ, G.R., Sale, P.F., Steneck, R.S., Oppen, M.J.H. & Willis, B.L. (2009). Larval retention and connectivity among populations of corals and reef fishes: history, advances and challenges. *Coral Reefs*, **28**, 307–325.

Jones, G.P., Planes, S. & Thorrold, S.R. (2005b). Coral Reef Fish Larvae Settle Close to Home. *Current Biology*, **15**, 1314–1318.

Laitila, J. & Moilanen, A. (2013). Approximating the dispersal of multi-species ecological entities such as communities, ecosystems or habitat types. *Ecological Modelling*, **259**, 24–29.

Largier, J.L. (2003). Considerations in estimating larval dispersal distances from oceanographic data. *Ecological Applications*.

Lehtomaki, J., Tomppo, E., Kuokkanen, P., Hanski, I. & Moilanen, A. (2009). Applying spatial conservation prioritization software and high-resolution GIS data to a national-scale study in forest conservation. *Forest Ecology and Management*, **258**, 2439–2449.

Levin, S.A., Muller-Landau, H.C. & Nathan, R. (2003). The ecology and evolution of seed dispersal: a theoretical perspective. *Annual Review of Ecology*.

Moilanen, A. & Wintle, B.A. (2006). Uncertainty analysis favours selection of spatially aggregated reserve networks. *Biological Conservation*, **129**, 427–434.

Moore, J.L. & RUNGE, M.C. (2012). Combining Structured Decision Making and Value-of-Information Analyses to Identify Robust Management Strategies. *Conservation Biology*, **26**, 810–820.

Moran, E.V. & Clark, J.S. (2012). Between-Site Differences in the Scale of Dispersal and Gene Flow in Red Oak (B. Fenton, Ed.). *PLoS ONE*, **7**, e36492.

Nathan, R. (2006). Long-Distance Dispersal of Plants. *Science (New York, N.Y.)*, **313**, 786–788.

Nathan, R., Klein, E., Robledo-Arnuncio, J.J. & Revilla, E. (2012). Dispersal kernels: review. *Dispersal Ecology and Evolution* (eds J. Clobert, M. Baguette, T.G. Benton & J.M. Bullock), p. 496. Oxford University Press, Oxford.

Neubert, M.G. & Parker, I.M. (2004). Projecting Rates of Spread for Invasive Species. *Risk Analysis*, **24**, 817–831.

Okubo, A. & Levin, S.A. (2013). *Diffusion and ecological problems: modern perspectives*.

Planes, S., Jones, G. & Thorrold, S. (2009). Larval dispersal connects fish populations in a network of marine protected areas. *Proceedings of the National Academy of Sciences*, **106**, 5693–5700.

Runge, M.C., Converse, S.J. & Lyons, J.E. (2011). Which uncertainty? Using expert elicitation and expected value of information to design an adaptive program. *Biological Conservation*.

Ruttenberg, B.I., Haupt, A.J., Chiriboga, A.I. & Warner, R.R. (2005). Patterns, causes and consequences of regional variation in the ecology and life history of a reef fish. *Oecologia*, **145**, 394–403.

Salomon, Y., Connolly, S.R. & Bode, L. (2010). Effects of asymmetric dispersal on the coexistence of competing species. *Ecology Letters*, **13**, 432–441.

Skellam, J.G. (1951). Random Dispersal in Theoretical Populations. *Biometrika*, **38**, 196.

Sponaugle, S. & Grorud-Colvert, K. (2006). Temperature-mediated variation in early life history traits and recruitment success of the coral reef fish Thalassoma bifasciatum in the Florida Keys. *Marine Ecology Progress Series*.

Telfer, S., Piertney, S.B., Dallas, J.F., Stewart, W.A., Marshall, F., Gow, J.L. & Lambin, X. (2003). Parentage assignment detects frequent and large-scale dispersal in water voles. *Molecular Ecology*, **12**, 1939–1949.

Wang, J. (2014). Estimation of migration rates from marker-based parentage analysis. *Molecular Ecology*.

Warner, R.R. (1997). Evolutionary ecology: how to reconcile pelagic dispersal with local adaptation. *Coral Reefs*, **16**, S115–S120.

White, C., Kendall, B.E., Gaines, S., Siegel, D.A. & Costello, C. (2008). Marine reserve effects on fishery profit. *Ecol Lett*, **11**, 370–379.

White, J.W. (2015). Marine reserve design theory for species with ontogenetic migration. *Biology Letters*, **11**, 20140511–20140511.

Woltmann, S., Sherry, T.W. & Kreiser, B.R. (2012). A genetic approach to estimating natal dispersal distances and self-recruitment in resident rainforest birds. *Journal of avian biology*, **43**, 33–42.

FIGURES AND TABLES

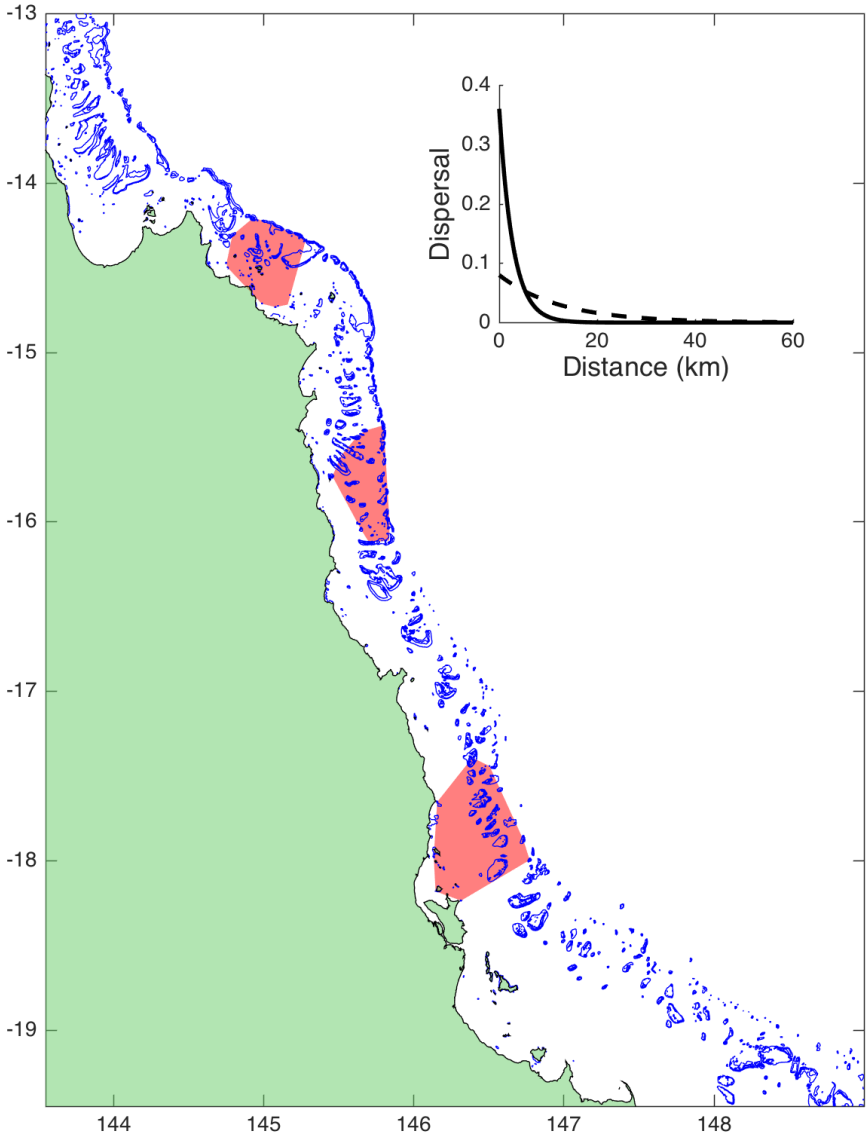


Figure 1: Location of the reefs on the Great Barrier Reef, Australia, used for the power analysis simulations. Red polygons demarcate sets of 40 reefs used to generate parentage data and fit dispersal kernels in three bootstrap runs. Inset panel shows the scale of the two dispersal kernels used for the power analyses.

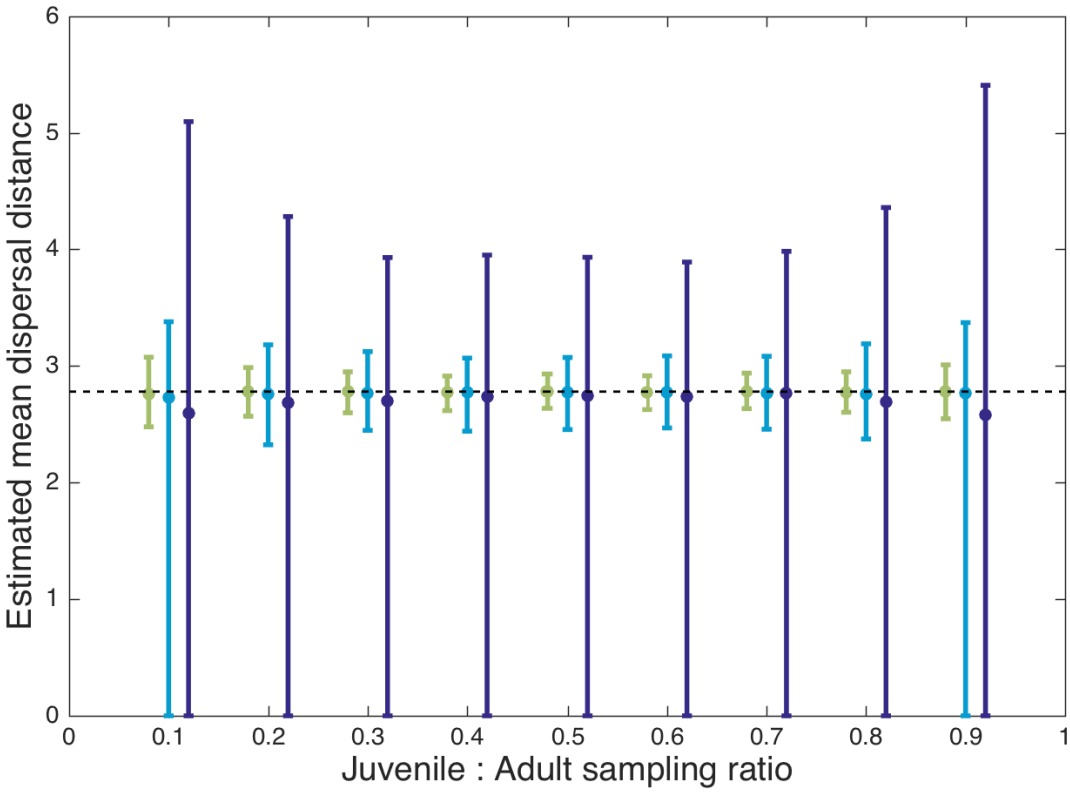


Figure 2: Accuracy of the kernel estimator for different allocations of sampling effort between adults and juveniles. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the short-distance dispersal kernel (long-distance results are in Supplementary Figure S1), and for 1,000 (purple), 2,500 (blue) and 5,000 juveniles (green), sampled across 25 reefs. Best results are achieved when sampling is equally distributed between adults and juveniles.

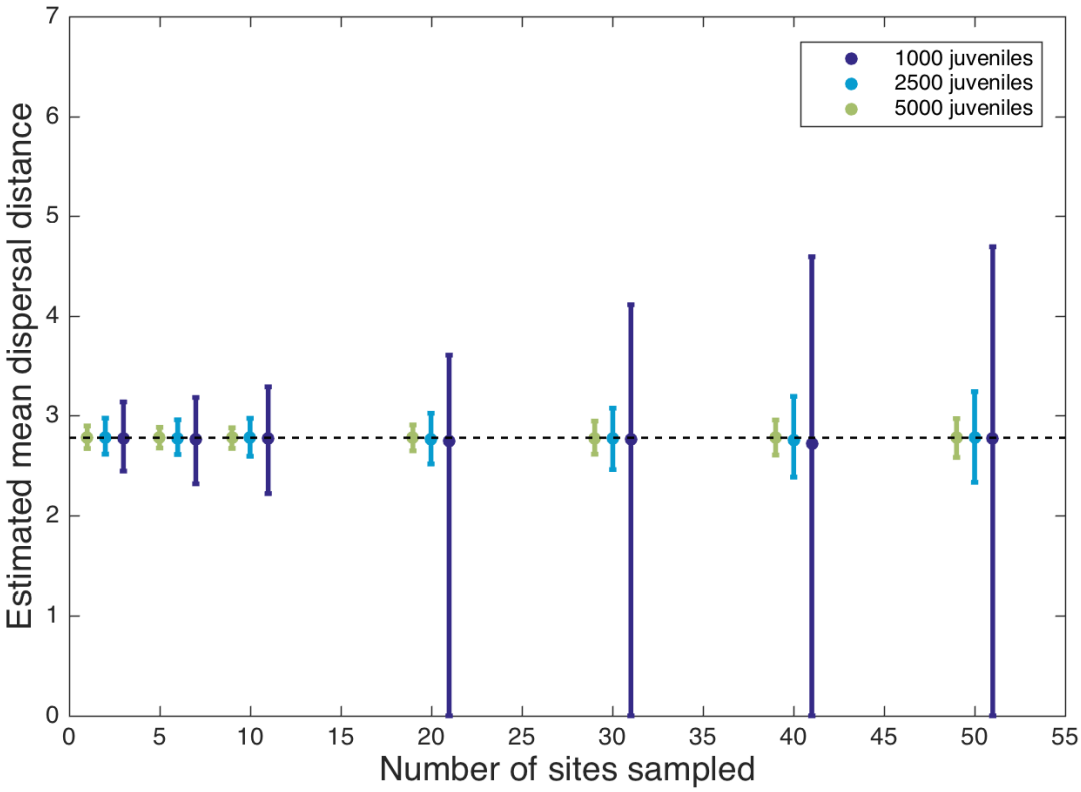


Figure 3: Accuracy of the kernel estimator, for different numbers of sampled reefs. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the short-distance dispersal kernel (long-distance results are in Supplementary Figure S2) for three levels of total sampling effort (colors), distributed equally between adults and juveniles. Sampling a wider range of sites degrades the performance of the estimator.

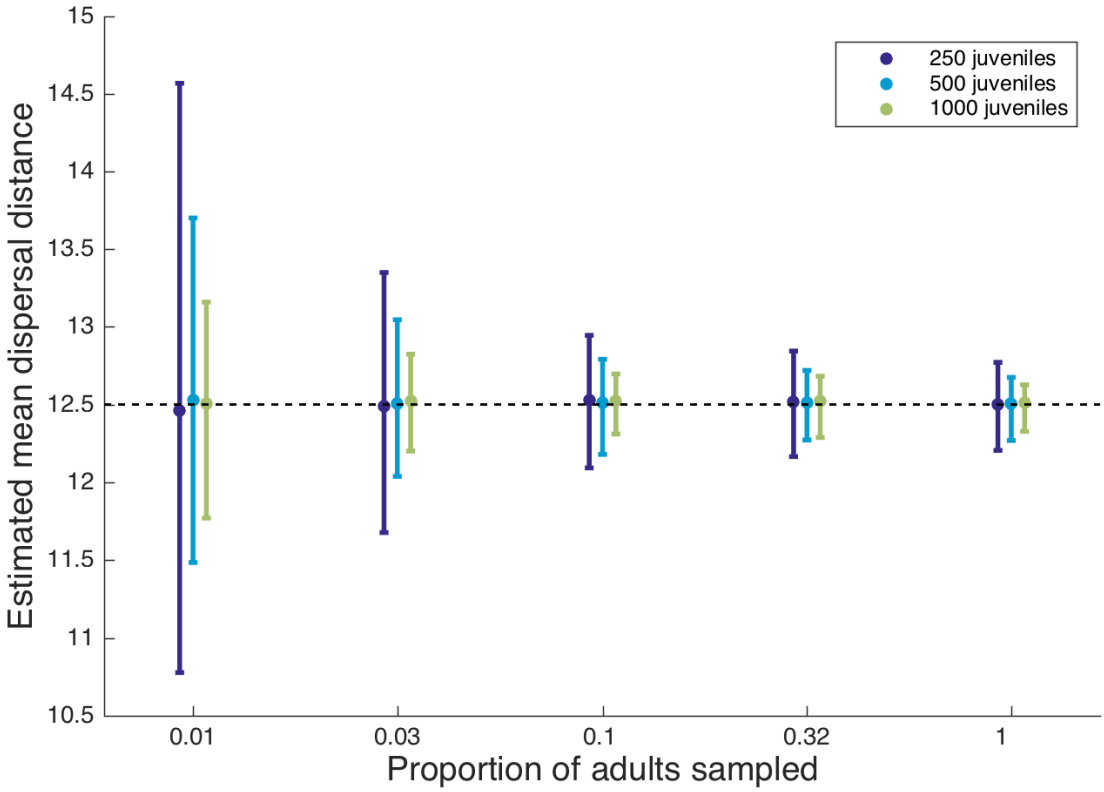


Figure 4: Accuracy of the kernel estimator, for proportions of adults sampled on each patch. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the long-distance dispersal kernel (short-distance results are in Supplementary Figure S3), and three levels of total sampling effort distributed across 25 reefs. Results are very poor when the proportion falls below 5%, but are relatively unchanged above this proportion.

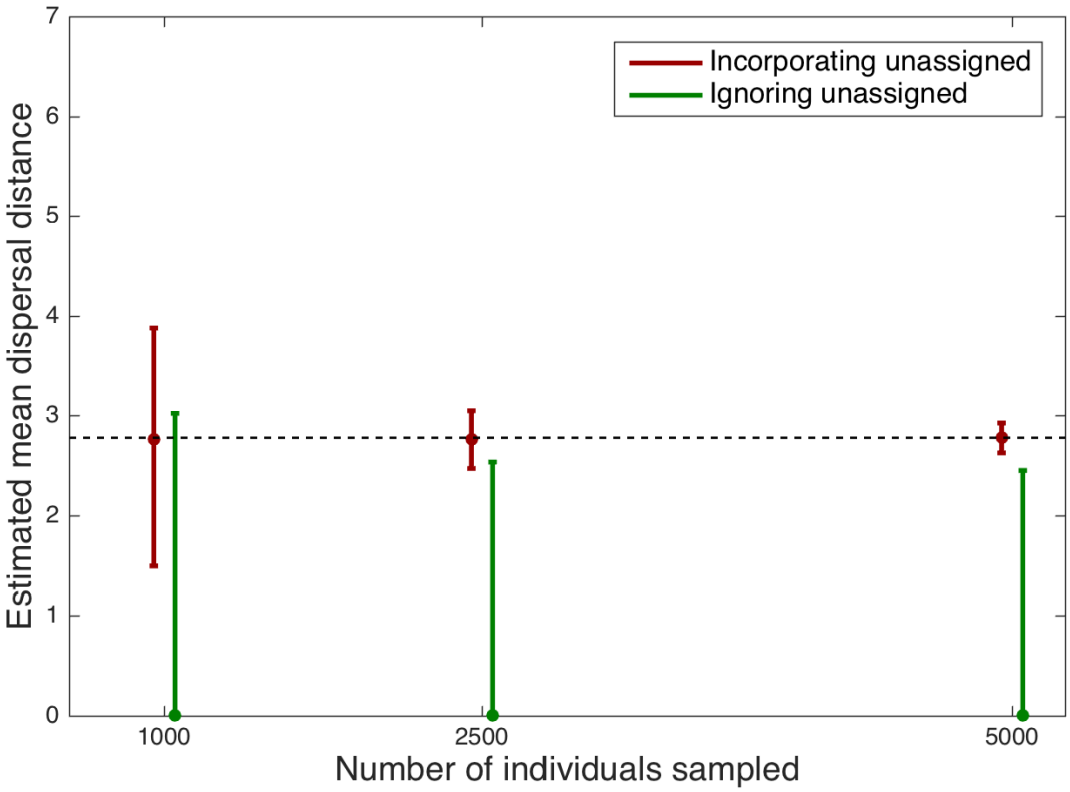


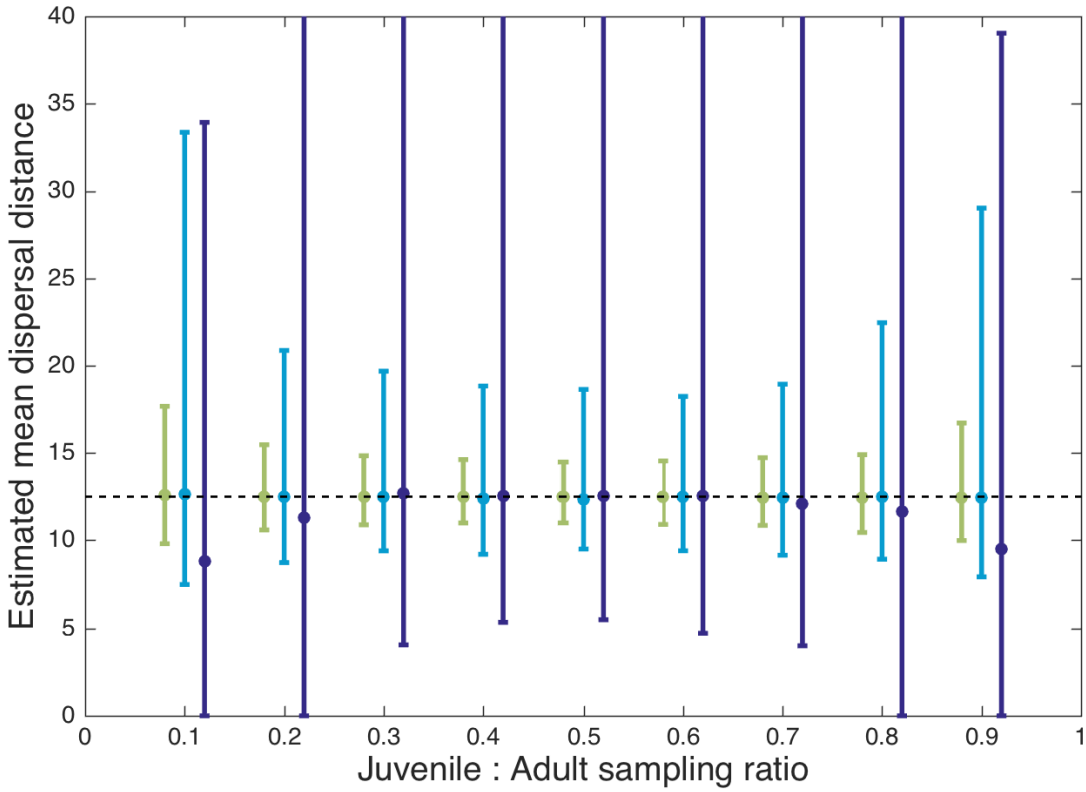
Figure 5: Accuracy of the kernel estimator when ghost populations are accounted for (red), and when they are ignored (green). Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for three levels of total sampling effort, distributed evenly between adults and juveniles across 25 reefs. (Long-distance results are in Supplementary Figure S4)

		Juvenile population		
Adult population		A	B	C
	A	5	0	0
	B	20	60	20
	C	1	3	3
	X	11	13	9

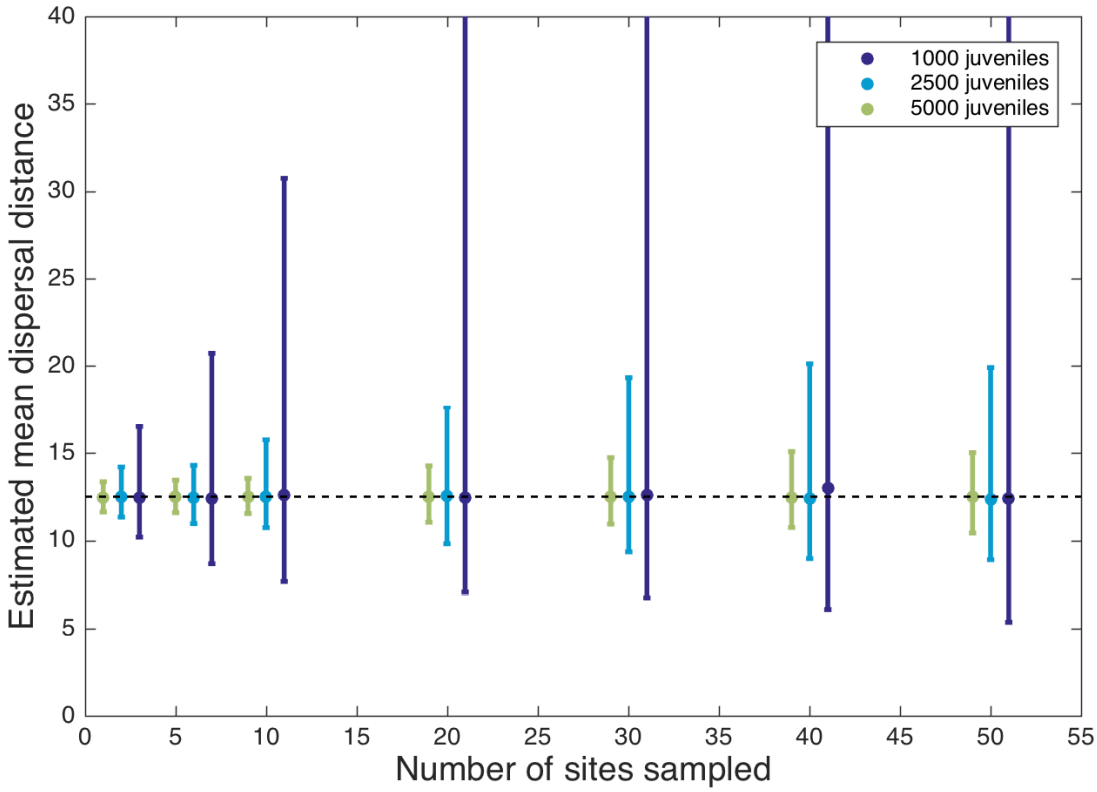
Table 1: Example parentage matrix **M**. Note that the sampled matrix is highly asymmetric, and that certain rows have large values, reflecting large source populations. The final row (labelled X) corresponds to sampled juveniles that could not be assigned to a sampled adult.

Parentage analysis study	Sampling scale (approx.)	Adults sampled		Juveniles sampled		Best-fit larval dispersal kernel		
		Total #	Sites	Total #	Sites	Kernel function	Mean dist	ln(k)
P. maculatus Keppels Island Group, Australia	30 km	466 (30%)	19	493 (58) (12%)	19	Negative exponential	12.5 km	-2.5
Elacatinus lori Carrie Bow Cay, Belize	41 km	3,033 (3%)	30	4,122 (120) (3%)	30	Negative exponential	2.8 km	-1.0
P. areolatus Manus Island, Papua New Guinea	55 km	416 (43%)	1	782 (76) (10%)	66	Ribbens		-10.1

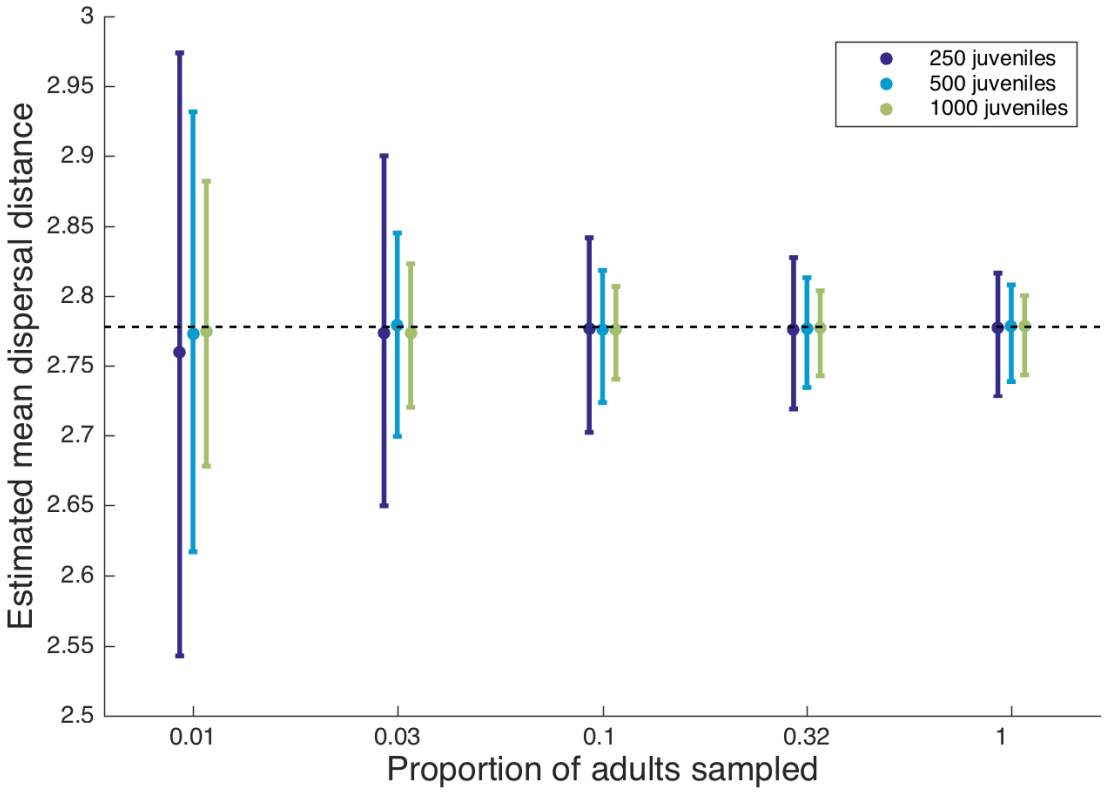
Table 2: Details of the three case studies used to ground the exploration of parameter space. All values are taken from the references or their supplementary information. Case studies are from Harrison *et al.* (2012), Almany *et al.* (2013), and D'Aloia *et al.* (2015).



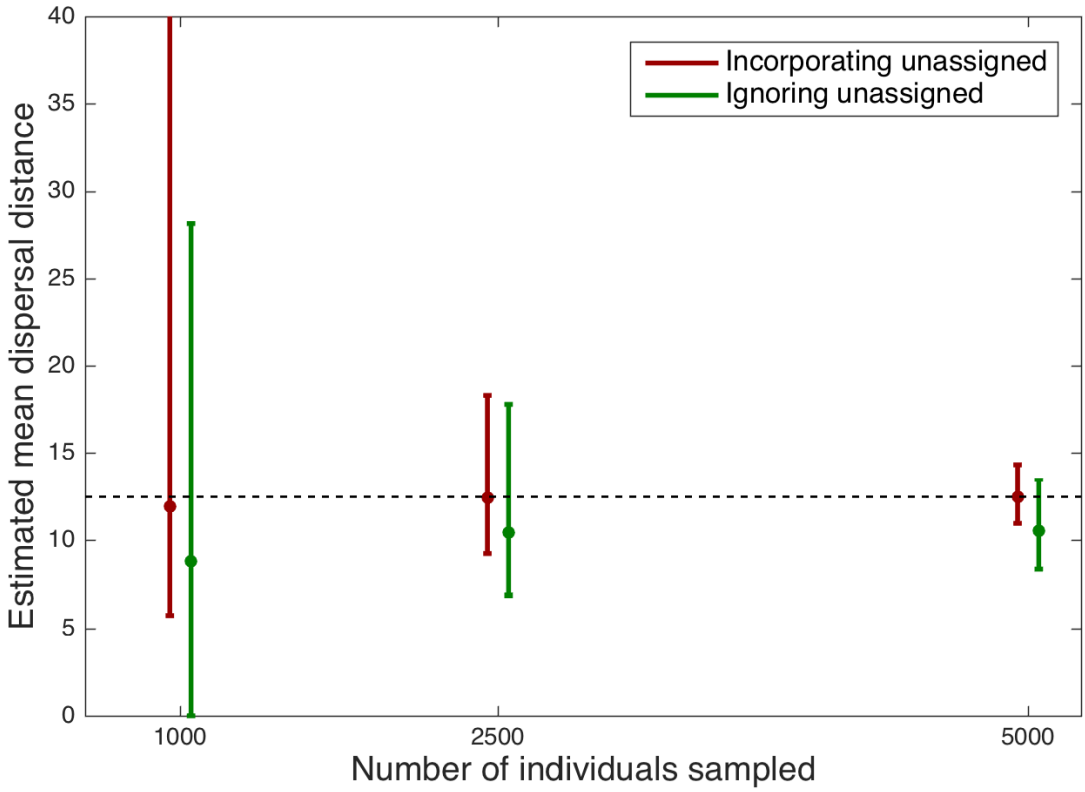
Supplementary Figure S1: Accuracy of the kernel estimator for different allocations of sampling effort between adults and juveniles. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the long-distance dispersal kernel, and for 1,000 (purple), 2,500 (blue) and 5,000 juveniles (green), sampled across 25 reefs. Best results are achieved when sampling is equally distributed between adults and juveniles.



Supplementary Figure S2: Accuracy of the kernel estimator, for different numbers of sampled reefs. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the long-distance dispersal kernel for three levels of total sampling effort (colors), distributed equally between adults and juveniles. Sampling a wider range of sites degrades the performance of the estimator.



Supplementary Figure S3: Accuracy of the kernel estimator, for proportions of adults sampled on each patch. Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the short-distance dispersal kernel, and three levels of total sampling effort distributed across 25 reefs. Results are very poor when the proportion falls below 5%, but are relatively unchanged above this proportion.



Supplementary Figure S4: Accuracy of the kernel estimator when ghost populations are accounted for (red), and when they are ignored (green). Dashed line shows the true mean dispersal distance for the species; circles show the estimated mean dispersal distance based on fits to the parentage data; error bars enclose 95% of 1,000 datasets, sampled from across the region. Results are shown for the long-distance disperser, for three levels of total sampling effort, distributed evenly between adults and juveniles across 25 reefs.