Statistical Science
2016, Vol. 31, No. 2, 259–274
DOI: 10.1214/16-STS552

Close-Kin Mark-Recapture

Mark V. Bravington, Hans J. Skaug and Eric C. Anderson

Abstract. Mark-recapture (MR) methods are commonly used to study wildlife populations. Taking advantage of modern genetics one can generalize from "recapture of self" to "recapture of closely-related kin". Abundance and other demographic parameters of adults can then be estimated using, if necessary, only samples from dead animals (live-release is optional). This greatly widens the scope of MR, e.g. to commercial fisheries where large-scale tagging is impractical, and enhances the power of conventional MR studies where live release and tissue sampling is possible. We give explicit formulae for kinship (i.e., recapture) probabilities in general and specific cases. These yield a pseudo-likelihood based on pairwise comparisons of individuals in the samples. It is shown that the pseudo-likelihood approximates the full likelihood under sparse sampling of large populations. Experimental design is addressed via the principle of maximizing the Fisher information for parameters of interest. Finally, we discuss challenges related to kinship determination from genetic data, focusing on current limitations and future possibilities.

Key words and phrases: Demography, genetics, kinship, mark-recapture, pseudo-likelihood.

1. INTRODUCTION

Mark-recapture (MR) methods allow population size and demographic parameters such as mortality rates to be estimated from an appropriately designed marking experiment. The modern foundation was laid in Cormack (1964), Jolly (1965) and Seber (1965), and in the subsequent 50 years MR methods have been widely used in wildlife studies, as well as in other fields such as sociology. The basic premise is that individuals can be marked by some means, and that the marks will be recognized if the individual is recaptured at some later sampling occasion. In the present paper, we focus on DNA marks, which are becoming increasingly popular in wildlife studies (Blouin, 2003, Lukacs and Burnham, 2005). A major advantage of DNA tags is that they contain additional information about relatedness among individuals in the sample. Based on this observation, Skaug (2001) suggested a "single sample" version of the Petersen mark-recapture estimator. An individual is marked by its presence in the sample, and "recaptured" if the sample contains one or more close relatives; intuitively, this is less likely to happen in bigger populations, so the number of "recaptures" provides information on adult abundance. The recaptures also provide information on other demographic parameters: for example, adult survival rate is inversely related to the average time taken to "recapture" a parent after sampling its offspring. We refer to methods that use information about relatedness in the sample as close-kin mark-recapture (CKMR), and in this paper we extend the classical MR framework by incorporating close-kin information. Bravington (2015) established an abundance estimate for adult Southern bluefin tuna based on the detection of 45 parent-offspring pairs in 13,000 samples, constituting the first large scale application of CKMR.

To clarify the ideas, the simplest version of CKMR is shown in Figure 1. Each juvenile is an offspring which

Mark V. Bravington is a Statistician, CSIRO Marine Lab, GPO Box 1538, Hobart 7001, TAS, Australia (e-mail: Mark.Bravington@csiro.au). Hans J. Skaug is Professor, Department of Mathematics, University of Bergen, P.O. Box 7803, N-5020 Bergen, Norway (e-mail: skaug@math.uib.no). Eric C. Anderson is Research Molecular Geneticist, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 110 Shaffer Road, Santa Cruz, California 95060, USA (e-mail: Eric.Anderson@noaa.gov).

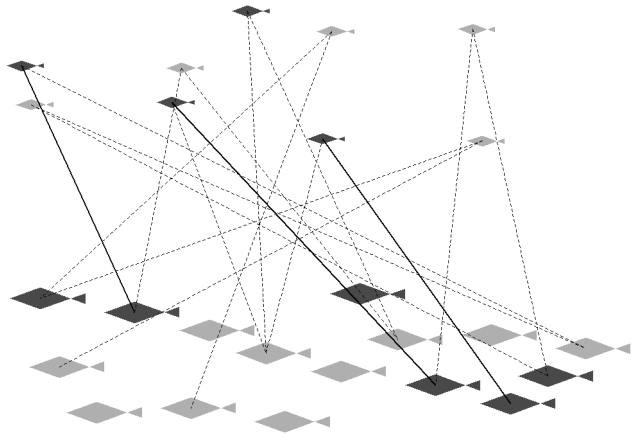

FIG. 1. *The simplest form of CKMR. Juveniles are small, adults are big; parents and offspring are linked by lines; dark means sampled, light means unsampled.*

"marks" its two adult parents. We compare the genotype of each of the n_J juvenile samples to each of the n_A adult samples, to check if a "mark" is recaptured. The probability that the adult happens to be one of the juvenile's two parents is $2/N_A$, where N_A is adult population size. Hence, if the entire set of $n_J \times n_A$ comparisons yields H Parent-Offspring pairs, then adult abundance can be estimated as $\hat{N}_A = 2n_J n_A/H$. Real applications to open populations are more complicated due *inter alia* to adult mortality in the interval between birth and sampling, non-random sampling, reproductive variability, different types of "mark" (i.e., different kinships) and uncertainty in genotyping, all of which can affect the probability of recapture.

CK greatly expands the scope of MR, most obviously because all samples can be obtained merely from dead animals if necessary, thus avoiding the often expensive or intractable task of live-release. Hence, CKMR is directly and cheaply applicable to "fully lethal" settings such as hunting, by-catch and large-scale commercial fisheries. Much of this paper concentrates on large-population settings, which we anticipate will be the most important arena for CKMR.

From a statistical perspective, extending MR to CK involves three main challenges:

1. Different demographic equations and extra parameters (e.g., fecundity-at-age) needed to describe "recapture" probabilities;
2. A more complicated likelihood because the latent space of possible histories is much more intricate (in effect, the genealogy of the entire population);
3. Quantifying the reliability of kinship-determination from genotype data (loosely analogous to misidentification in photo-ID studies).

We address these points in the rest of this paper, first presenting motivating examples and setting up the core assumptions in Section 2. In Section 3, we develop generic equations for prior probability of parent-offspring and half-sibling pairs in terms of individual-level covariates, and show examples for specific scenarios. Section 4 shows how the prior probabilities can be combined into a pseudo-likelihood, similar to Skaug (2001). It also considers the link to Fisher information, and suggests simple approaches to experimental design, including the crucial question of sample size. We consider genotyping in Section 5, reviewing how kin can be identified statistically, and considering which types of kinship can reliably be determined with genetic technology available now (parent-offspring, half-sibling) and in the future. In the discussion of Section 6, we mention the (many) strengths and (few) weaknesses of CKMR, and discuss statistical challenges for future work. The whole field is quite new, being made possible only by technological advances from the mid-2000s in cost and reliability of genotyping, so this paper is as much a preview as a review. The technological progress shows no sign of slowing down, so the scope of CKMR applications is likely to become wider and wider.

2. CLOSE-KIN METHODS

2.1 Motivating Examples

CKMR methods are more varied than standard MR in the sense that the actual implementation depends considerably on the biology of the species in question, especially the reproductive biology. Our own experience of CKMR currently comprises 7 or 8 projects on bony fish (teleosts), marine and riverine sharks, and marine mammals; project status ranges from desk designs to completed exercises. To illustrate something of the variety and scope, and to provide points-of-reference for the statistical developments later on, we briefly describe three real examples: Southern Bluefin Tuna, White Sharks and Antarctic Blue Whales. Although none are terrestrial, they do vary considerably with respect to biology and with respect to the covariate information that is available. They illustrate two clear differences between CKMR and MR: that fecundity is important, and that time-of-marking is unknown unless age is measured. See Table 1 for abbreviations of kinship categories.

TABLE 1

Kinship categories (K) with their identity-by-descent (ibd) *probabilities κ (see Section 5). Abbreviation codes are given for the kinship types that occur frequently in this paper. For asymmetric kinships, such as parent-offspring, a distinction in the notation (PO versus OP) is made to indicate which individual is the parent. The unrelated category (U) will often include all kinships not explicitly mentioned in a CKMR model, in which case the* ibd *probabilities shown will only be approximate*

Pairwise kinship	Abbrev. K	*ibd* probability κ_0	κ_1	κ_2
Self, Monozygous twin	S	0	0	1
Parent–offspring	PO/OP	0	1	0
Mother–offspring	MO/OM			
Father–offspring	FO/OF			
Full sibling		1/4	1/2	1/4
Half-sibling	HS	1/2	1/2	0
Maternal half-sibling	MHS	1/2	1/2	0
Grandparent–grandchild		1/2	1/2	0
Aunt–niece		1/2	1/2	0
Unrelated	U	1	0	0

Southern bluefin tuna (*Thunnus maccoyi*). This valuable, long-lived fish has been heavily exploited for decades, with the adult stock reduced to perhaps 5% of pre-exploitation biomass by the mid 2000s (Hillary et al., 2015). While slow recovery is now expected under improved management, considerable uncertainty remains both about the demography and about the reliability of underlying data, especially for monitoring adult abundance. To address this, sampling for CKMR began in 2006, with 3-year-old juveniles from a nursery-ground fishery, and adults from the spawning-ground fishery where age, sex and length were also recorded. By 2012, about 8000 juveniles and 5000 adults had been genotyped at 20–25 hypervariable microsatellite loci, and 45 PO pairs were found among the approximately 38,000,000 adult-juvenile comparisons. An age-length-sex-structured CKMR model was built to estimate adult abundance (as a time series), fecundity-at-size and adult mortality rate (Bravington, Grewe and Davies, 2014). The results—indicating an adult stock just under 2,000,000 animals—have been adopted into management, and CKMR is being extended to long-term monitoring of the species. In the future, new genotyping technologies (Section 5.3) will enable juvenile half-sibling pairs to be identified; this improves the information per sample, and also allows relaxation of a biological assumption required by the PO-only model.

White sharks (*Carcharodon carcharias*). This is a late-maturing animal with modest litter sizes, little growth after maturity and presumed low resilience to exploitation. The Eastern Australia/New Zealand population was exposed to appreciable human-induced mortality in the middle of the last century. There is some public perception that numbers may have increased recently, but no suitable data for constructing any kind of abundance estimate, nor for reliably monitoring trends in the future. Adult white sharks are too rarely encountered to be useful in any abundance-estimation method, but juveniles can be reliably sampled along the eastern Australian coast. A CKMR study began in 2011, using tissue samples from living and dead juveniles to identify half- (and full-) siblings, with age estimated from body length or vertebral ring counts. So far about 20 HS pairs have been found, and a preliminary CKMR model has been fitted to estimate abundance (again, a time-series) and survival.

Antarctic blue whales (*Balaenoptera musculus intermedia*). This subspecies was almost extirpated by whaling between 1900 and the early 1970s. It is unclear how well the population has since recovered; subsequent abundance estimates from line transect surveys are in the low thousands, but with high CVs because of low encounter rates. However, with a long-term mark-recapture project using biopsy sampling (DNA tags), it may be possible to acoustically locate enough whales to make a respectable abundance and trend estimate using classical MR methods (Peel et al., 2013). If the population is indeed recovering as expected, then PO pairs would increase the number of "recaptures per sample" substantially, and a design study has shown that CVs would be considerably reduced (Bravington, Grewe and Davies, 2014). Hence, this would constitute a CKMR application with non-lethal sampling only. The lack of age information from biopsies (though see Section 6) complicates the calculation of kinship probabilities (Section 3.5).

2.2 General Framework

In the standard MR literature, the data from each individual are organized into capture histories (Pollock et al., 1990). In a close-kin setting, this is difficult for a number of reasons, such as linked capture histories of related individuals and potentially because of uncertainty in kin identification. Instead, we consider a pseudo-likelihood approach which uses only pairwise comparisons of sampled individuals. A pairwise comparison between individuals i and j gives rise to a kinship category K_{ij} taking values from a discrete set $\mathcal{K}$:

TABLE 2

Summary of notation (excluding genetics; see Section 5). In the third column "latent" denotes quantities that may be unobserved, but are needed in model formulation. Note that N and n refer to different quantities, and hence do not conform to the standard upper/lower case letter convention used in statistics

Term	Definition	Status
n	Sample size (# captured individuals)	Observed
p	Capture probability at given sampling occasion	
	Quantities related to a captured individual	
y	Time (year) of birth	Latent
x	Place of birth or capture	Latent/Observed
t	Time of capture	Observed
z	Covariate vector at time of capture	Observed
a	Age	Latent/Observed
ℓ	Length	Observed
	Population dynamics and demography	
N	Total population size (males and females)	Parameter
ϕ	Individual survival probability	Parameter
α	Age of maturity	Parameter
β	Per capita average birth rate, $\beta = \mathbb{E}(R)$	Parameter
R	Reproductive output (# offspring) of an individual in a given year	Latent
R_+	Total reproductive output from all individuals in a given year	Latent
	General	
θ	Vector of all model parameters	Parameter
K	Kinship category (see Table 1)	Latent/Observed

for example, $\mathcal{K} = \{\text{PO}, \text{HS}, \text{U}\}$, where the categories are explained in Table 1. Depending on the situation, only certain categories may be relevant: for example, with lethal sampling, self recapture (S) is impossible; if only juveniles are caught, then parent–offspring (PO) is impossible; if too few genetic markers are used, half-siblings (HS) may be undetectable. The unrelated category (U) covers all more distant relationships than those explicitly modelled.

2.2.1 *Population dynamics and demography.* To emphasize the relationship to classical MR, we have chosen our notation (Table 2) to match that literature as closely as possible. We treat time t as a discretely varying quantity, and for simplicity we refer to the unit of time as a "year". There are three broad classes of parameters that determine the population dynamics for CKMR:

ϕ Individual survival probability, which in general can vary with age and other covariates.

β Per capita birth rate, i.e. the expected number of offspring an individual produces in a given year. This parameter may also vary with age and other covariates. The actual reproductive output of a given individual in year t is denoted by $R(t)$, and is a random variable with $\mathbb{E}[R(t)] = \beta$.

N_t Adult population size in year t. Given the population size in some reference year t_0, the subsequent population trajectory will evolve deterministically or stochastically according to ϕ and β.

Subscripts are used when the population is stratified; for example, the number of females aged a in year t is denoted by $N_{\varphi at}$. We do not go into much detail about the actual form of the population dynamics, as the CKMR approach does not require any special form. Complex population dynamics are allowed for, such as density dependence where both ϕ and β may be functions of N_t, and thereby vary with time. N_t itself can be treated either as deterministic or random, with demographic stochasticity handled as latent random variables on survival and/or incoming recruitments; in our motivating examples, we ignored stochastic survival (a reasonable approximation for large populations) and we allowed for stochastic recruitment only in the case of tuna, which like many bony fish shows large fluctuations in year-class strength. At the individual level, on the other hand, demographic stochasticity certainly plays an important role.

Sampling of individuals is performed at discrete time points. The sampling may be with or without removal from the population. A typical example of the former

is lethal "sampling" which is relevant for catch data. Non-lethal sampling introduces the minor notational problem that individuals i and j in the sample actually may refer to the same biological individual. Associated with each sampled individual there may be covariates, such as date and location of capture, age and sex. As noted above, the demographic parameters such as ϕ and β may depend on some of these covariates. Similarly, the sampling probability p of a given individual in the population may depend on its covariates. This allows for stratified sampling, e.g. if sampling is done only among juvenile individuals $[p_t(\text{adult}) = 0]$ for certain t, as for southern bluefin tuna. We shall denote by z the vector of observed covariates. Further assumptions on the sampling scheme are given in Section 3.

Throughout much of this manuscript, we assume that sampling is sparse, i.e. only a small fraction of the population is sampled, by which we mean that the total sample size is $n = O(N^{1/2})$, where N is adult population size. As will be shown in Section 4.3, this yields $O(1)$ number of recaptures, and thus $O(1)$ precision on estimated parameters. Sparse sampling is not a requirement of close-kin methods in general, but when sampling is not sparse the pseudo-likelihood presented below may not have the properties expected of a proper likelihood.

3. KINSHIP PROBABILITIES

We now show how the kinship probability $\mathbb{P}(K_{ij} = k|z_i, z_j)$ can be calculated for parent–offspring ($k =$ PO) and half-sibling ($k =$ HS) relationships. These probabilities are the building blocks of the pseudo-likelihood defined later. Our notation indicates that we are conditioning on the covariate vectors z_i and z_j, meaning that we take a stochastic view of the covariates. Recall that z_i consists of all the information about an individual, except its genotype and kinship to other individuals, that is recorded when the individual is captured. We include t_i in z_i, as we will always condition on the time and fact of capture. A second subscript, or an argument, on a covariate always refers to time, and indicates the (unobserved) value of the covariate at a time other than sampling; thus, ℓ_i would be i's length when sampled at t_i, and ℓ_{it^*} or $\ell_i(t^*)$ would be i's length at some other time t^*. Note that there may also be important covariates that are not measured and thus not in z, e.g. age in Section 3.1.3.

It is in general necessary to deal separately with maternal and paternal descent, and the sexes may have quite different demographics. For simplicity, we concentrate here on mother–offspring (MO) and maternal half-sibling (MHS) kin only, and we assume adult sex is known; extensions to other cases are straightforward.

3.1 Mother–Offspring

Consider a female i, an offspring j, and the event "i is j's mother", i.e. $K_{ij} = \text{MO}$. The key to calculating $\mathbb{P}[K_{ij} = \text{MO}|z_i, z_j]$ is the latent variable $R_i(x_j, y_j)$ or just R_{ij} for short, namely i's reproductive output at the place x_j and time y_j that j was born. Given R, the probability that j is i's offspring is just R_{ij} divided by $R_+(x_j, y_j)$, the total reproductive output of all adult females at that same time and place. Note that x_j and y_j may or may not be uniquely determined by z_j, and that place of birth is irrelevant unless there is some kind of heritable stock structure; if place does matter, then the notation allows i to reproduce in more than one place per year.

The above assumes that z is irrelevant once R is known, i.e. $\mathbb{P}[K_{ij} = \text{MO}|R, z] = \mathbb{P}[K_{ij} = \text{MO}|R]$. For this to hold, "reproductive output" should strictly speaking be measured in the correct units, namely the number of surviving "juveniles" at t_j that have $z = z_j$ and that therefore have the same sampling probability as j; hence, R could in principle also have z_j as a third argument. However, it will often be reasonable to assume that all offspring from the same (x, y) have equal survival and sampling probabilities, and to work with a general notion of "fecundity" β (sometimes as a function of age or other parental covariates) as the measure of expected reproductive output. For simplicity, we make that assumption throughout, though it can be relaxed. Two further conditions on covariates are given at the end of Section 3.1.1.

The notion of kinship probability for MO as expected relative reproductive output, can then be expressed directly albeit opaquely as

$$\begin{aligned}\mathbb{P}[K_{ij} = \text{MO}|z_i, z_j] \\ = \mathbb{E}_{Y_j, X_j|z_j}\left[\frac{\mathbb{E}[R_i(X_j, Y_j)|z_i]}{\mathbb{E}[R_+(X_j, Y_j)]}\right].\end{aligned} \tag{3.1}$$

Practical use entails re-expressing this in terms of case-specific covariates and parameters. We show some examples below, starting with simple cases for clarity, and noting connections with our motivating examples.

3.1.1 *Base case: Ages known, constant fecundity, lethal sampling, no stock structure.* Since date-of-capture and age are known, i.e. $z = (t, a)$, we also know dates of birth $y = t - a$. Clearly, i can be j's

mother if and only if i was alive and mature at y_j (and not yet sampled, since sampling is lethal). In that case, thanks to constant fecundity she would be just as likely as any of the other $N_{♀y_j}$ mature females to be j's mother, so her expected relative reproductive output would be $1/N_{♀y_j}$. Writing α for age-at-maturity and $\mathbb{I}[\cdot]$ for the indicator function, the composite result is

$$\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] = \frac{\mathbb{E}[R_i(y_j)|y_{i,}, t_i]}{\mathbb{E}[R_+(y_j)]} = \frac{\mathbb{I}[y_i + \alpha \le y_j < t_i]}{N_{♀y_j}}. \tag{3.2}$$

This expression is unaffected by year-to-year fluctuations in the overall survival probability of juvenile cohorts, because we condition on the fact that j is alive when sampled. For a similar reason, adult survival probabilities do not appear explicitly, although they do indirectly affect $N_{♀y_j}$. Adult sampling probabilities p do not appear either. There is no requirement that all adults should have the same sampling probability, nor indeed the same expected reproductive output (though the latter should not change systematically after maturity: Section 3.1.3).

What does matter, however, is that the event of an adult's being sampled should be independent of the number of its offspring sampled, conditional on covariates. As well as z_i and z_j, there may also be relevant unobserved covariates, which then need to be integrated over as in the examples below. This key requirement can be expressed formally as two linked conditions:

1. For the potential mother i, there are no additional covariates that affect *both* the probability of her being sampled (p) *and* her expected reproductive output at y_j (R_{ij}). This is analogous to requiring "no unmodelled heterogeneity of capture probability" in conventional MR; its failure here would mean that some mothers are more likely to be "marked" because they have more offspring, and also more (or less) likely to be "recaptured". A counterexample is suggested at the end of Section 3.1.4.
2. There are no additional covariates that affect the sampling and/or survival probabilities of *both* i and her offspring. One counterexample would be if blueness is heritable, and blue animals are more catchable, but colour is mistakenly omitted from the conditioning. Another would be if mother-calf pairs are caught together.

3.1.2 *Non-lethal sampling*. If animals are not killed by sampling (e.g., if biopsies are used), it is no longer necessary that $t_i < y_j$, because i can continue to produce offspring after being sampled. However, she might still die naturally between t_i and y_j, so her expected reproductive output at y_j must be discounted accordingly. The kinship probability becomes

$$\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] = \frac{\mathbb{I}[y_i + \alpha \le y_j]}{N_{♀y_j}} \times \begin{cases} 1; & t_i > y_j, \\ \phi_i(t_i, y_j); & t_i < y_j, \end{cases} \tag{3.3}$$

where $\phi_i(t_i, y_j)$ is the probability that i survives from t_i to y_j.

3.1.3 *Age-dependent sampling and fecundity; ages unknown*. We now introduce age-dependent sampling probability $p(a)$, and age-specific fecundity $\beta(a)$. Sampling probability does not affect either the lethal or non-lethal case when age is known. However, age-specific fecundity does affect reproductive output. For example, the lethal case (3.2) becomes

$$\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] = \frac{\mathbb{I}[y_j < t_i] \times \beta(a_i - (t_i - y_j))}{\sum_{a \ge \alpha} \beta(a) N_{♀a y_j}}, \tag{3.4}$$

where the maturity constraint $y_i + \alpha \le y_j$ has been absorbed into the argument of $\beta(\cdot)$. By definition, $\beta(a) = 0$ for $a < \alpha$, including the case $a < 0$ to avoid notational awkwardness in the numerator.

We first extend this so that a_j is still observed but a_i is not, by summing the numerator over the possible ages of i at t_i. Her age distribution is conditional on the fact that she was sampled, so $p(a)$ now becomes relevant:

$$\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] = \frac{\mathbb{I}[y_j < t_i]}{\sum_{a \ge \alpha} \beta(a) N_{♀a y_j}} \times \frac{\sum_{a>0} \{p(a) N_{♀a t_j} \times \beta(a - (t_i - y_j))\}}{\sum_{a>0} p(a) N_{♀a t_j}}. \tag{3.5}$$

If a_j is also unknown, then so is j's year-of-birth y_j, and a further level of summation is required, again involving $p(a)$ but this time for j. We omit the formula for brevity.

In our blue whale design study, we considered non-lethal variants of this case, with $p(a)$ and $\beta(a)$ constant for $a \ge \alpha$ and zero otherwise, and simple deterministic demography leading to closed-form expressions for the sums.

3.1.4 *Covariates other than age.* Age, whether measured or not, is crucial to CKMR because it determines the time of "marking". However, it may not be the sole covariate of importance. For example, in many teleost fish including tuna, growth continues after reaching age of maturity, and fecundity and sampling probability ("selectivity") are primarily related to body length rather than age. Assuming both age and length are measured, writing $\beta(\ell)$ for fecundity-at-length, and letting the density of length-at-age at y_j be $f(\ell|a, y_j)$, then the total reproductive output at y_j is now a weighted sum: $\sum_{a\geq\alpha} N_{\female a y_j} \int \beta(\ell) f(\ell|a, y_j)\, d\ell$. Provided i was alive and mature at y_j, her expected reproductive output then would be determined by her length, $\ell_i(y_j)$. Although we do not observe $\ell_i(y_j)$ directly, we can hindcast it from her measurements at capture, $z_i = \{\ell_i, a_i, t_i\}$, through an individual-specific growth curve. This largely matches our tuna example. Under lethal sampling again, the kinship probability is

$$\begin{aligned}\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] &= \frac{\mathbb{I}[y_i + \alpha \leq y_j < t_i] \times \beta(\ell_i(y_j))}{\sum_{a\geq\alpha} N_{\female a y_j} \int \beta(\ell) f(\ell|a, y_j)\, d\ell}.\end{aligned} \tag{3.6}$$

Note that ignoring the dependence on length, e.g. by naively applying (3.2) instead, would violate the first assumption in Section 3.1.1. If big fish are more likely to breed as well as to be caught, this resembles "heterogeneity of capture probability" in Jolly–Seber models which, if overlooked, is a well-known source of bias in abundance estimates (Pledger, Pollock and Norris, 2010).

3.1.5 *Stock structure.* Suppose stocks are associated with discrete locations $\mathcal{X}$, and that x_{it} or $x_i(t)$ denotes i's location in year t so that $x_j(y_j)$ is j's stock-of-birth, and that animals are sampled later somewhere in $\mathcal{X}$ though not necessarily in their original stock. If sampling location is known, then the relative locations within kin-pairs carry considerable information about stock structure (Kanno, Vokoun and Letcher, 2011, Palsbøll, Peery and Berube, 2010). The modification to kinship probability is easy to state: we need to sum over the possible locations where j might have been born, taking account of the probability that i was there at the right time. For example, in the base case (3.2) we have

$$\begin{aligned}\mathbb{P}[K_{ij} = \mathrm{MO}|z_i, z_j] &= \sum_{x\in\mathcal{X}} \Big\{ \mathbb{P}[X_{jy_j} = x|z_{jt_j}]\mathbb{P}[X_{iy_j} = x|z_{it_i}] \\ &\quad \times \frac{\mathbb{I}[y_i + \alpha \leq y_j < t_i]}{N_{\female x y_j}} \Big\},\end{aligned} \tag{3.7}$$

where $N_{\female xt}$ is now the number of female adults in x at time t. A comprehensive treatment would require a movement model between location-of-birth (and of breeding) and location-of-sampling, perhaps akin to Spatially Explicit Capture-Recapture (Borchers and Efford, 2008). The biological, sampling and statistical possibilities of CKMR with stock structure are still to be explored.

3.2 Half-Siblings

Two samples i and j are maternal half-sibs (MHS) if and only if they share the same mother and have different fathers. For simplicity, we assume that ages are known, that without loss of generality i was born first (i.e., $y_i < y_j$), and that mating is random in a large population, so that full-sibs (at least from different breeding occasions) are negligibly rare. Since we have no direct observations on i or j's mother(s), we need to sum across all possible mothers d in the set $\mathcal{F}_i$ of adult females alive at i's year-of-birth. Ignoring stock structure and other partly-heritable covariates, we have

$$\begin{aligned}\mathbb{P}[K_{ij} = \mathrm{MHS}|y_i, y_j] &= \sum_{d\in\mathcal{F}_i} \{\mathbb{P}[i\text{'s mother was } d|y_i] \\ &\quad \times \mathbb{P}[j\text{'s mother was } d|y_j]\}.\end{aligned} \tag{3.8}$$

Again, the key is latent variables: the probability that d is i's mother is proportional to her (unobserved) reproductive output $R_d(y_i)$ when i was born. Again writing $R_+(y)$ for the total reproductive output at y, (3.8) can be re-expressed as

$$\begin{aligned}\mathbb{P}[K_{ij} = \mathrm{MHS}|y_i, y_j, R] &= \sum_{d\in\mathcal{F}_i} \left\{ \frac{R_d(y_i)}{R_+(y_i)} \times \frac{R_d(y_j)}{R_+(y_j)} \right\},\end{aligned} \tag{3.9}$$

where $R_d(y_j) = 0$ if d dies between y_i and y_j. In practice, we need expectations of $R_d(y)$, which are also determined by fecundity $\beta_d(y)$, and thus may vary with d and y. Although it is possible to write an explicit and general formulation of (3.9) incorporating fecundity and mortality, the end result is unenlightening, and in practice it may be more useful to head for a case-specific formula. For example, in a base case where adult fecundity and mortality rates are constant across age and time, we obtain

$$\begin{aligned}\mathbb{P}[K_{ij} = \mathrm{MHS}|y_i, y_j] = N_{\female y_i} \times \left\{ \frac{1}{N_{\female y_i}} \times \frac{\phi(y_i, y_j)}{N_{\female y_j}} \right\} \\ = \frac{\phi(y_i, y_j)}{N_{\female y_j}},\end{aligned} \tag{3.10}$$

where summation over the $N_{♀y_i}$ identical females in $\mathcal{F}_i$ has become a multiplication. This is basically what we have used for white sharks, where we can only sample juveniles, and female growth after maturity is limited and litter sizes do not seem strongly linked to female body size.

If $y_i = y_j$ in (3.9), then the two multiplicands refer to the same random variables, and the overall expectation depends not just on the mean across individuals of R/R_+ but also its variance. This is generally unknown, but can be very large in animals with big litters and high early-life-stage mortality, such as many teleost fish. Problems can be avoided by restricting HS comparisons to different cohorts, i.e. where $y_i < y_j$ strictly.

There are subtle differences between the HS and PO cases. For instance, (3.9) assumes that actual (as opposed to expected) reproductive output between years is uncorrelated. Unmodelled heterogeneity in fecundity—i.e. consistent over time, variable across individuals, but *not* modelled—would lead to bias in $\mathbb{P}[K_{ij} = \text{HS}]$ and in the corresponding abundance estimate; as an extreme example, infertile adults are invisible to an HS-only analysis. With PO CKMR, such heterogeneity only matters if it is somehow correlated with sampling probability. Also, with HS it seems even more important to measure age, at least approximately, partly because of the same-cohort issue above, partly because the ability to estimate mortality rates in, for example, (3.10) depends on knowing y, and partly because age can be the only way to distinguish HS pairs from the genetically-similar grandparent–grandchild pairs (Section 5).

Sibship is also connected with short-term "effective population size" (N_e: Crow and Kimura, 1970), and indeed Wang (2009) uses within-cohort sibship to estimate N_e. The difference between N_e and true (census) adult population size is driven by overall variance in individual reproductive output; our approach here attempts to model or exclude as much as possible of that variance.

4. PSEUDO LIKELIHOOD AND DESIGN

Let θ denote the vector of all parameters that govern the basic quantities N, ϕ, β and p. In this section, we deal with estimation of θ, parameter identifiability, and issues related to design of experiments. We assume for now that all pairwise kinships K_{ij} can be ascertained with certainty from the genetics; Section 5 deals with uncertain kinship.

4.1 Pseudo-Likelihood

The joint distribution of $\{K_{ij}; 1 \le i < j \le n\}$ is too complicated to permit constructing a full likelihood. Instead, we use a pseudo-likelihood approach (Besag, 1975, Skaug, 2001) which involves only the marginal probabilities of the K_{ij}:

$$l_P(\theta) = \sum_{1 \le i < j \le n} \log \mathbb{P}(K_{ij} = k_{ij} | z_i, z_j; \theta), \tag{4.1}$$

where k_{ij} denotes the observed kinship, and $\mathbb{P}(K_{ij} = k_{ij}|z_i, z_j; \theta)$ is one of the probabilities derived in Section 3. Each term in (4.1) conditions on z_i and z_j, but clearly z itself may contain information about some parts of θ (e.g., age composition is highly informative about mortality). To account for this, one may add the marginal log likelihood contribution from the z to obtain an alternative pseudo-likelihood

$$l_P(\theta) = \sum_{1 \le i < j \le n} \log \mathbb{P}(K_{ij} = k_{ij} | z_i, z_j; \theta) + \sum_{i=1}^{n} \log \mathbb{P}(z_i; \theta). \tag{4.2}$$

Both (4.1) and (4.2) give rise to unbiased estimating equations, in the sense that the pseudo-score has zero expectation; i.e.

$$\mathbb{E}_{K,Z}\left[\frac{\partial}{\partial \theta} l_P(\theta; K, Z)\Big|_{\theta=\theta_{\text{true}}}\right] = 0, \tag{4.3}$$

where θ_{true} denotes the true parameter values under which the expectation is taken. This unbiasedness is the condition required for proving consistency of the maximum pseudo-likelihood estimator $\hat{\theta}(k, z)$ satisfying $[\partial l_P(k, z)/\partial \theta]_{\hat{\theta}} = 0$. The choice between (4.1) and (4.2) is case-specific. One might generally expect the marginal distribution of z to be informative, though as noted below the pseudo-likelihood may not always approximate the true joint likelihood of all K_{ij}, in which case simply adding the two terms as in (4.2) may not be optimal. Sometimes use of (4.2) is essential for identifiability, for example, with length frequency data in our tuna example. Such issues are beyond the scope of the present paper.

4.2 Parameter Identifiability and Experimental Design

The Fisher information matrix I can be used to study parameter identifiability, as Cole et al. (2012) have done in MR. More generally, I quantifies the amount of information about each individual parameter, and

hence constitutes a key tool when designing a mark-recapture experiment. In this section, we argue that the per-pair Fisher information, i.e. that of each individual term in the pseudo-likelihood (4.1), provides a valuable tool for certain purposes. Working for simplicity with (4.1) rather than (4.2), the conditional per-pair Fisher information is

$$\begin{aligned} & I(\theta|z_i, z_j) \\ & = -\mathbb{E}_K\left[\frac{\partial^2}{\partial\theta^2}\log\mathbb{P}(K|z_i, z_i; \theta)\right] \\ & = -\sum_{k\in\mathcal{K}}\frac{\partial^2\log\{\mathbb{P}(k|z_i, z_j; \theta)\}}{\partial\theta^2}\mathbb{P}(k|z_i, z_j; \theta). \end{aligned}$$

To quantify the information content in a typical observation pair, it is necessary to also take expectation with respect to covariates,

$$I(\theta) = -\mathbb{E}_{Z_i,Z_j}[I(\theta; Z_i, Z_j)]. \tag{4.4}$$

The latter expectation can be evaluated empirically. Expression (4.4) is useful for comparing (relative) precision across different experimental designs. Two examples where relative precision might suffice are: comparison of different sets of kinship $\mathcal{K}$ (to see whether extra genotyping effort is worthwhile); and allocation of sampling effort across different occasions.

If one seeks absolute precision of estimators, more elaborate calculations may be called for. However, the next section gives conditions under which the pseudo-likelihood is approximately the full likelihood, in which case (4.4) (together with any information from marginal likelihoods of z) is adequate for absolute precision, too.

4.3 Large Populations, Sparse Samples, and Crude Designs

In a short-term study (less than a typical adult lifespan, say), each pairwise comparison has probability $O(N^{-1})$ of yielding a kin pair, where N denotes average adult abundance. With a sample size of n, there are $O(n^2)$ comparisons (the precise number depends on how many kin-types are considered, and on whether some comparisons are pointless because covariates z rule out the possibility of kinship). If H is the number of kin-pairs found, then its expected value h_{nN} is $cn^2N^{-1} + o(n^2N^{-1})$ for some c which depends on the design and the demographics, but not on n or N. This suggests setting $n_{\text{target}} \approx \sqrt{Nh_{\text{target}}/c} \sim O(N^{1/2})$ to achieve a reasonably informative dataset with, say, h_{target} in the range 50–100. For moderately large N, H is the sum of many Bernoulli outcomes each with low expected value, and is approximately Poisson provided the comparisons are approximately independent (see below). Thus, an approximate lower bound on the CV (coefficient of variation) of $\hat{N}$ is $1/\sqrt{h_{\text{target}}}$, e.g. 15% for $h_{\text{target}} = 50$. The real CV may of course differ since N itself may be time-dependent, comparisons are not strictly independent, and other parameters also must be estimated, but the lower bound can still be useful for a crude design. For example, in the single-sample setting of Figure 1 and with an equal mix of juvenile and adult samples, $c = 1/2$ so a 15% CV requires $n \approx 10\sqrt{N}$. In fact, we have found $10\sqrt{N}$ to be a useful reference point for feasibility in several short-term PO-only applications, but the multiplier 10 is certainly not universal; any serious design should at least evaluate h_{nN} by summing over the expected number of comparisons and the likely probabilities of kinship.

As to approximate independence: the set of pairwise comparisons cannot be fully independent, because one outcome can sometimes predict another. If j's mother is already found, she cannot be found again. However, a heuristic justification can be made provided sampling is sparse enough for expected kin-triads (within which the pairwise comparisons are clearly not independent) to be rare compared to kin-pairs. If $n \approx O(N^{-1/2})$, as per the previous paragraph, then the ratio of triads to pairs is $O_p(N^{-1/2})$, so sparse sampling amounts to N being large enough. (In longer-term studies with substantial turnover of adults, total n can presumably be larger, because triads will remain rare.) The key point is that only a small proportion of samples will be involved in any kin-pair at all. For most comparisons between one sample i and another j, there will be little predictive power even from knowing the outcomes of all other comparisons involving i and j, since all those outcomes will usually just be Unrelated, and thus largely uninformative about i's relationship to j. In other words, under sparse sampling the expected marginal information from a comparison is similar to its expected conditional information given all other comparisons, so approximate independence is reasonable.

One further caveat on approximate independence is that the samples should not contain a high proportion of "littermates" (full- or half-sibs in the same cohort), since comparisons between two littermates and any third animal are obviously not independent. This could be violated even in large sparsely-sampled populations, if litters are large and show persistent schooling behaviour through to the time of a school-based sampling

process (e.g., larval sampling of some teleost fish). If this is a realistic concern, then one might need to check the incidence of within-cohort half-sibs among the samples. Provided the K_{ij} are sufficiently independent, it is easy to show that the pseudo-likelihood obeys not just the score property (4.4) exactly, but also approximately the key second-derivative property,

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l_P(\theta)\Big|_{\theta=\theta_{\text{true}}}\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} l_P\right)\left(\frac{\partial}{\partial\theta} l_P\right)^{\top}\Big|_{\theta=\theta_{\text{true}}}\right], \tag{4.5}$$

which ensures that it "scales" in the way a true likelihood should. These properties are the keystones of likelihood-based inference, so if (4.5) is satisfied, then it is reasonable to treat the pseudo-likelihood as a true likelihood, for instance, when incorporating other data such as the sampling distributions of z.

If there is concern about the adequacy of (4.5) in a particular setting, i.e. whether N is big enough and n small enough for "sparsity" to hold, then it can be checked by individual-based (not pair-based) simulation using a guess at θ_{true}. For each simulated sample of individuals, l_P is evaluated, and its first- and second-order derivatives are obtained using numerical or automatic differentiation (Griewank and Walther, 2008). The expected values on both sides of (4.5) are evaluated as averages across simulations. If the approximation turns out to be poor, then one might consider other approaches to estimating equation inference, such as the "sandwich method" (Huber, 1967). Small-population CKMR in general is a topic for further work.

5. GENETIC IDENTIFICATION OF CLOSE-KIN PAIRS

The previous sections have assumed that the kinship k_{ij} between each pair of samples is exactly known; however, k_{ij} must be inferred from genetic data. In practice, for a variety of reasons, k_{ij} may not be perfectly known or knowable. For example, the genetic data available may be insufficient to correctly identify every kin pair without incurring false positives among the unrelated pairs; or, certain kin categories may be indistinguishable given the genetic data, such as OP versus PO, or HS versus grandparent–grandchild. With an adequate level of DNA data, however, this uncertainty is surmountable and can be handled within the pseudo-likelihood framework. In the next subsection, we review statistical genetic background for developing a likelihood-ratio kin identification statistic and computing false positive and negative rates. Following that, we explain how kinship uncertainty can be accommodated in the pseudo-likelihood, either via classification with allowance for error, or via treating true kinship as a latent variable. Finally, we conclude with a short appraisal of the future prospects for CKMR in today's genomic era.

5.1 Inference of Kin Pairs from Genotype Data

The genetic data that can be observed on an individual are called its genotype. For a diploid organism, the genotype g is a random vector of discrete-valued pairs

$$g = (g_1, \ldots, g_L) = \big((g_{1a}, g_{1b}), (g_{2a}, g_{2b}), \ldots, (g_{La}, g_{Lb})\big).$$

Each pair $g_l = (g_{la}, g_{lb})$ corresponds to a single location ("locus" or "marker") in the nuclear genome and the two values (g_{la}, g_{lb}) refer to the two copies of each marker carried by the individual. Although one copy is inherited from the mother and the other from the father, the parental origin of each copy is seldom known, so the two values within each pair are taken to be unordered. The values g_{la} or g_{lb} denote the specific, detectable variants ("alleles" or "allelic types") at the marker. Multiple alleles may be present in the population, though any individual will carry at most two of them. Genotypes are often assumed drawn from a population in Hardy–Weinberg equilibrium and without linkage disequilibrium (LD: Weir, 1996), in which case the marginal probability of an individual i's genotype is

$$\mathbb{P}[g^{(i)}] = \prod_{l=1}^{L} \mathbb{P}[g_l^{(i)}] = \prod_{l=1}^{L} \pi(g_{la}^{(i)})\pi(g_{lb}^{(i)})(1 + \mathbb{I}[g_{la}^{(i)} \neq g_{lb}^{(i)}]),$$

where $\pi(g_{la}^{(i)})$ is the frequency in the population of the allele carried at the ath gene copy at the lth locus within individual i and $\mathbb{I}[g_{la}^{(i)} \neq g_{lb}^{(i)}]$ is an indicator function, taking 1 if $g_{la}^{(i)} = g_{lb}^{(i)}$ and 0 otherwise.

Inference of k_{ij} relies on the joint probability $\mathbb{P}[g^{(i)}, g^{(j)}|k_{ij}]$, which is influenced by k_{ij} because relatives can share genes that are either direct descendants of one another or descendants of the same gene that existed in a recent common ancestor. Such genes are said to be "identical-by-descent" or *ibd* (Thompson, 2013).

Since gene mutation rates are low enough to be considered negligible over short time scales, if two gene copies at a locus within different individuals are *ibd* they will be the same allele, and if they are not *ibd* the allelic types are independent of one another. Thus, relatives of a certain type k_{ij} have genotypes that are similar in predictable ways depending on the fraction of genome that they share *ibd*. At two non-inbred individuals (i.e., where the gene copies at a locus within either individual are not *ibd*), at a single locus l, the joint probability $\mathbb{P}[g_l^{(i)}, g_l^{(j)}|k_{ij}]$ can be entirely specified by the expected fraction of loci at which the pair shares 0, 1, or 2 gene copies *ibd*, denoted κ_0, κ_1, and κ_2, respectively, with $\kappa_0 + \kappa_1 + \kappa_2 = 1$ (Cotterman, 1940). Values of $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ for a few common kinships (k_{ij}'s) appear in Table 1. At locus l, we have

$$\begin{aligned}
&\mathbb{P}[g_l^{(i)}, g_l^{(j)}|k_{ij}] \\
&\quad = \kappa_0 \mathbb{P}[g_l^{(i)}]\mathbb{P}[g_l^{(j)}] \\
&\qquad + \kappa_1 \mathbb{P}[g_l^{(i)}, g_l^{(j)}|k_{ij} = \mathrm{PO}] \\
&\qquad + \kappa_2 \mathbb{P}[g_l^{(i)}]\mathbb{I}[g_l^{(i)} = g_l^{(j)}],
\end{aligned} \tag{5.1}$$

where $\mathbb{P}[g_l^{(i)}, g_l^{(j)}|k_{ij} = \mathrm{PO}]$ is the probability of i and j's genotypes at locus l, given that i and j share exactly 1 gene copy *ibd* (as they would if they were a parent–offspring pair). The value of $\mathbb{P}[g_l^{(i)}, g_l^{(j)}|k_{ij} = \mathrm{PO}]$ is easily determined from laws of Mendelian inheritance (Table 3). If the L markers of the multilocus genotype are physically unlinked (on different chromosomes or chromosome arms) and not in LD, then (5.1) extends easily:

$$\mathbb{P}[g^{(i)}, g^{(j)}|k_{ij}] = \prod_{l=1}^{L} \mathbb{P}[g_l^{(i)}, g_l^{(i)}|k_{ij}]. \tag{5.2}$$

Moreover, in the case of pairs with kinship S, PO/OP, or U (Table 1), the factorization in (5.2) holds even if the markers are physically linked, so long as they are not in LD (Anderson and Garza, 2006). It is worth noting that the relationship categories of half-siblings, grandparent–grandchild, and aunt–niece share the same value of k, and accordingly cannot be distinguished with physically unlinked markers.

TABLE 3

Probabilities of the genotype of individual j given that it is an offspring of individual i. A_a, A_b, and A_c denote distinct allelic types and π_a [short for $\pi(a)$], π_b, and π_c are the relative frequencies of those alleles in the population

	Genotype of j			
Genotype of i	A_aA_a	A_aA_b	A_aA_c	A_bA_c
A_aA_a	π_a	π_b	π_c	0
A_aA_b	$\frac{1}{2}\pi_a$	$\frac{1}{2}(\pi_a + \pi_b)$	$\frac{1}{2}\pi_c$	$\frac{1}{2}\pi_c$

Thompson (1975) used (5.2) to develop a general likelihood framework for estimating κ and inferring k_{ij} of a pair of individuals. Applying that framework to CKMR requires a modification to allow for genotyping errors. A number of models have been proposed to account for genotyping errors in relationship estimation (Marshall et al., 1998, Sobel, Papp and Lange, 2002, Wang, 2004), and models accounting for allelic dropout (failure to score an allele present in an individual) have been used in mark-recapture of individuals identified from non-invasive genetic samples (Wright et al., 2009, Barker et al., 2014). However, more generally, *any* single-locus model for genotyping error can be specified in terms of $\mathbb{P}[g_l^*|g_l]$, the probabilities of the recorded genotype g_l^* given the true genotype g_l, and then accommodated by using

$$\begin{aligned}
&\mathbb{P}[g^{(i)*}, g^{(j)*}|k_{ij}] \\
&\quad = \sum_{g^{(i)}, g^{(j)}} \{\mathbb{P}[g^{(i)*}|g^{(i)}] \times \mathbb{P}[g^{(j)*}|g^{(j)}] \\
&\qquad \times \mathbb{P}[g^{(i)}, g^{(j)}|k_{ij}]\},
\end{aligned} \tag{5.3}$$

which is easily computed when (5.2) factorizes over loci, in place of $\mathbb{P}[g^{(i)}, g^{(j)}|k_{ij}]$. The genotyping error rate for each locus can be approximately estimated by the genotyping concordance rate for multiply-genotyped individuals, but it is better estimated by using genotypes from related pairs of individuals (Wang, 2010).

In CKMR, related pairs are sought from amongst a large number of unrelated pairs, requiring a method for statistical error control to minimize the rate at which unrelated pairs are falsely classified as related (Skaug, Berube and Palsbøll, 2010). In natural populations (Meagher and Thompson, 1986), the classification of pairs to kinship categories is typically done via the likelihood ratio

$$\Lambda_k(i, j) = \frac{\mathbb{P}[g^{(i)*}, g^{(j)*}|K_{ij} = k]}{\mathbb{P}[g^{(i)*}]\mathbb{P}[g^{(j)*}]}. \tag{5.4}$$

The distribution of $\Lambda_k(i, j)$ when i and j are truly unrelated, i.e. $\mathbb{P}[\Lambda_k < c|K_{ij} = \mathrm{U}]$ for any c, can be predicted from allele frequencies and genotyping error rates, so that the false positive rates expected from classifying according to a threshold value Λ_k^* can be

derived (Anderson and Garza, 2006). In large populations, the number of unrelated pairs will greatly exceed the number of true kin-pairs (by a factor of nearly 10^6 in our tuna study). Thus, it is essential to use enough informative markers to limit the overlap between the distributions of Λ_k for unrelated pairs and target kin-pairs. Suppressing false positives—a key part of the classification-with-error approach in the next section—may require setting Λ_k^* so high that some fraction ψ_k^- of true k-type kinpairs will have $\Lambda_k(i,j) < \Lambda_k^*$, thus becoming false negatives. ψ_k^- can be estimated in advance from the allele frequencies and genotyping error rates for any k if markers are unlinked, and for PO/OP and S pairs even if markers are linked but not in LD (Anderson and Garza, 2006). False negative rates for other kinships are harder, because more markers are needed and the distribution of Λ_k is then affected by physical linkage (see 5.3). However, provided enough definite pairs are present, then the observed distribution of $[\Lambda_k|\Lambda_k > \Lambda_k^*]$ may still be used to estimate ψ_k^-; this work is ongoing. As long as ψ_k^- can be estimated, it can be used analogously to allowing for failure to detect a mark in conventional MR, for example, by failing to match two photographs of the same animal in a photo-ID study. We summarize this process in the next subsection.

The foregoing has focused on nuclear autosomal markers, but two other types of loci with different modes of inheritance can also be useful for CKMR. Mitochondrial DNA (mtDNA) is a small ($\approx$ 16,000 base-pair), mostly non-recombining, DNA molecule that is maternally inherited; each individual has only one copy. Its allelic variants are called "haplotypes". Although mtDNA has low power for identifying kin pairs by its small size and the fact that it is inherited as a single locus, it can be useful afterward for inferring the sex of the parent in PO pairs, inferring the sex of the parent shared by HS pairs, or if sex is known, whether pairs are PO or OP. The Y chromosome, in organisms with XY sex-determination, is transmitted from the male parent to male offpring. Markers on the X and Y chromosomes can be useful for genetically determining the sex of samples and for identifying whether male half-sibling pairs have different fathers.

5.2 Kinship Uncertainty in the Pseudo-Likelihood

We have tried two ways to allow for kinship uncertainty: classification with allowance for error, and latent-kinship. In the former, we set the threshold for classification, Λ_k^*, so that the total number of expected false positives from U pairs is a small fraction (say, $<1\%$) of the number of observed pairs with $\Lambda_k > \Lambda_k^*$. We then estimate the false-negative probability ψ_k^-, and define a modified kinship-category k' based simply on comparing $\Lambda_k(i,j)$ to Λ_k^*. Since false-positives are negligible, we have

$$\begin{aligned}\mathbb{P}[K'_{ij} = k'|z_i, z_j; \theta] \\ = (1-\psi_k^-)\mathbb{P}[K_{ij} = k|z_i, z_j; \theta],\end{aligned}$$

so the pseudo-likelihood can easily be adjusted to k' rather than k. mtDNA can be handled similarly, defining modified kinships K'_{same} and K'_{diff} depending on whether i and j share a haplotype, and adjusting the prior probability of the modified kinships based on haplotype frequencies. In general, there can be several possible kinships for each pair (i,j), so we test sequentially from most- to least-related. Indistinguishable kinships, for example, PO as distinct from OP in the absence of age data (blue whales), are merged to a new category whose kinship probability is the sum of the constituent probabilities.

Classification has been effective for our tuna and shark examples; with good quality DNA and markers deliberately chosen to ensure good identifiability, few comparisons have any ambiguity and rates of false-negatives (where some information is discarded) have been below 5%. Classification also permits aggregation of comparisons into classes with equal kinship probabilities, as determined by (z_i, z_j); the number of pairs found in each class is a sum of independent Bernoulli trials, so it follows a Binomial distribution. In our tuna study, we could compress 38,000,000 comparisons into about 50,000 classes, leading to a 1000-fold speedup in subsequently calculating pseudo-likelihoods.

An alternative which makes full use of the genetic data, is to consider true kinship as a latent variable (Skaug, 2001). The response variable for each pair becomes all the genetic data (g_i, g_j), including mtDNA etc., with

$$\begin{aligned}\mathbb{P}[g_i, g_j|z_i, z_j; \theta] \\ = \sum_{k\in\mathcal{K}} \mathbb{P}[g_i, g_j|K_{ij} = k] \\ \times \mathbb{P}[K_{ij} = k|z_i, z_j; \theta],\end{aligned} \tag{5.5}$$

where $\mathcal{K}$ is the set of all possible kinships at whatever level of detail is necessary. The genetic probabilities are pre-computed as in this section, and the prior kinship probabilities are computed as in Section 3. Latent kinship avoids any loss of statistical efficiency through

false negatives, but computational efficiency may suffer since comparisons can no longer be aggregated. Latent kinship may be most valuable when DNA quality is poor, populations are small enough for appreciable inbreeding, and/or more distant and intrinsically hard-to-separate kinships are considered.

5.3 CKMR and Emerging Genetic Technologies

To date, mark-recapture using genetic markers has relied primarily on microsatellite markers (Wright et al., 2009, Barker et al., 2014, Bravington, Grewe and Davies, 2014), which are usefully multiallelic, but labor intensive to score and not easy to standardize between laboratories. However, within the last five years there has been a dramatic shift in the capacity to develop genotypes with many loci, even for species with previously-unstudied genomes (Davey et al., 2011). These gains have come largely from the ability to assay many SNP markers, leveraging microarray technology or recent advances in "next generation sequencing" (Shendure and Ji, 2008). SNPs usually have just two allelic types, and thus are individually less informative about kinship than microsatellites, but this is now outweighed by their far lower unit cost. We predict that SNP-based genotypes will become the predominant data type for CKMR. In our current experience, reliable identification of self-recaptures can be done with around 40 well-chosen SNPs, PO pairs with several hundred and HS pairs (subject to a small false negative rate) with one-to-a-few thousand, though these numbers do depend somewhat on population size and sample size. The genotyping cost per individual organism for any of those kinships can be reduced to tens of dollars if the best available technology is used.

The addition of many markers is not without some complications. Notably, when there are many markers per chromosome, it is no longer acceptable to assume that they are unlinked (i.e., inherited independently). Rather, nearby markers are inherited upon large, physically-linked chunks of genome, with two important consequences. First, there is a limit to how much power can be achieved by adding more markers, and thus a limit to how distantly-related the types of kinship used for CKMR can be; and second, because the state of being *ibd* is no longer independent between markers, the factorization of (5.1) across loci is not valid. In human genetics, algorithms are available to compute genotype probabilities of kin with dense marker data (Lander and Green, 1987), and these have been considered for kinship inference (Skare, Sheehan and Egeland, 2009). One advantage of using linked markers is that some kinship categories which have identical genotype probabilities with unlinked markers (half-sibling, grandparent–grandchild, aunt–niece) can be resolved with linked markers. However, accounting fully for physical linkage requires knowledge of the physical ordering of the markers along the chromosomes, and of the recombination rates between them.

Developing an accurate linkage map of a genome is still (in 2016) a costly and difficult undertaking, and it is unlikely that such maps will be developed quickly for many species to which CKMR might be applied. Fortunately, as already stated, for both unrelated and PO pairs, the joint probability (5.1) is correct regardless of physical linkage (at least in the absence of LD, which is usually reasonable if no more than a few thousand markers are used). For other kinships, we note that although exact calculation of $\mathbb{P}[g^{(i)*}, g^{(j)*}|k_{ij}]$ is unlikely to be possible without information on linkage, it is possible to compute $\Lambda_k(i, j)$ assuming the markers are unlinked. Conveniently, the distribution of Λ_k calculated thus is unaffected by physical linkage if the true relationship is Unrelated, so pairwise false positive rates from U-pairs can still be calculated for any threshold Λ_k^*, and false negative rates may be estimated empirically as mentioned earlier. In our experience, most half-siblings can be reliably inferred if enough SNPs are used, even without linkage information.

6. DISCUSSION

CKMR is a very new area made possible only by improvements in genetic technology in the last 5–10 years. There are many open methodological challenges, and the field may evolve considerably as CKMR is applied to more species. Even at this early stage, CKMR has a number of attractive features, including:

- Only needing samples from dead animals (though live biopsies are fine), so no expense or difficulty of live-release. This opens up applications in bycatch, roadkill, hunting and many commercial fisheries.
- Half-sibling methods permit the study of adults without ever catching them.
- No confounding from tag-reporting rate, since the presence/absence of a "tag" can be checked in and only in the lab (also true for standard MR with DNA-based marks). The experiment is hence "blinded" with respect to data collection.
- Less susceptibility to bias from "unmodelled heterogeneity of capture probability" because no self-recapture is involved and no recaptures are needed.

- CV is inverse to sample size not its square root (Section 4.3), so precision improves rapidly as samples accumulate. Often, samples can be collected cheaply, and only enough need be genotyped subsequently (the main expense) to obtain an informative number of kin-pairs.

The most fundamental limitation is that CKMR only informs directly about adults, regardless of whether juveniles are sampled. Of course, this may sometimes be resolved with other data and demographic modelling. The other limiting factor is the need to collect adequate covariates, adequate sample sizes, and adequate quality DNA—and to pay for genotyping which, while nowadays cheap (tens of US dollars per sample), is not free.

In addition, there are some species for which CKMR will presumably never work:

Parthenogenetic (virgin-birth, self-cloning) species such as whiptail lizards: for obvious reasons.

Armadillos (of genus *Dasypus*) which bear identical quadruplets: mothers are genetically indistinguishable from aunts.

Semelparous (breed-once-then-die) species that cannot be sampled as adults, such as eels and squid: inference is limited to successfully-breeding adults in previous generations, which is confounded with random reproductive variability (Section 3.2).

Some organisms, while not technically impossible, seem unlikely candidates for CKMR:

Super-abundant species, such as krill: the sample size n_{target} required to find a useful number of kin is economically daunting, even though n_{target} is proportional to $\sqrt{N}$ rather than to N itself (Section 4.3). As genotyping continues to drop in cost, though, more species will become affordable.

Very long-lived species which cannot be sampled young, such as orange roughy: since estimates are back-dated to juvenile birth, the results may not be useful.

Rarely-seen species, such as pygmy right whales: an impractically high fraction of the population might need sampling to yield a useful number of kin-pairs.

CKMR does require substantial biological insight, and sometimes extensive covariate data. Qualitatively, perhaps the ideal CKMR scenario is to sample adults across their lifespan as well as juveniles, to collect age/size and other relevant individual covariates, and to know enough about the basic biology to correctly assemble (3.1) and (3.9). Where this is impossible, less-ideal scenarios can often still work, but some populations may be ruled out through lack of data: e.g. essential covariates, background biological information or samples from important life-stages. The statistical manifestation is that a properly-formulated model becomes unidentifiable, but that a simplistic model may be biased (Skaug, 2001). For example, a juvenile-only entirely-HS study of teleost fish would not yield reliable adult abundance estimates because of confounding between abundance and lifetime changes in fecundity, though it could work for many mammals or sharks. Case-by-case consideration of (3.1) and (3.9), followed if necessary by examination of the Hessian for a trial design (Section 4.3), should provide the necessary insight; experimental design for CKMR is a topic of great practical importance requiring further work.

Age data is always helpful and in some cases essential; it can sometimes be obtained from otoliths, teeth or other hard-parts, and at other times an approximate inference from length might be adequate. In the future, though, age may come from the genetic samples themselves. Polanowski et al. (2014) successfully used epigenetic (DNA-methylation) signatures to estimate age in humpback whales, and advancing technology should improve accuracy (Jarman et al., 2015). In our blue whale design study, we showed that even low-accuracy epigenetic age data from biopsies, enough to tell which of a PO pair is the parent, would substantially improve precision.

We envisage that the main statistical challenges will be:

- Experimental design.
- Identification and use of more distant kin, as genotyping continues to improve.
- Extensions to spatial settings.
- Efficient adaptation to small populations and non-sparse sampling, where something closer to a full likelihood might be needed.

ACKNOWLEDGMENTS

This work was initiated when Hans J. Skaug and Mark Bravington were sabbatical visitors to the Center for Stock Assessment Research (CSTAR), a partnership between UCSC and NOAA Fisheries, Santa Cruz. Eric C. Anderson was supported by NSF-OCE-1260693, Mark Bravington by the Marine Biodiversity Hub of the Australian Government's NESP, and Hans J. Skaug by the Norwegian Research Council Grant 215867.

REFERENCES

ANDERSON, E. C. and GARZA, J. C. (2006). The power of single nucleotide polymorphisms for large-scale parentage inference. *Genetics* **172** 2567–2582.

BARKER, R. J., SCHOFIELD, M. R., WRIGHT, J. A., FRANTZ, A. C. and STEVENS, C. (2014). Closed-population capture-recapture modeling of samples drawn one at a time. *Biometrics* **70** 775–782. MR3295738

BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

BLOUIN, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.* **18** 503–511.

BORCHERS, D. L. and EFFORD, M. G. (2008). Spatially explicit maximum likelihood methods for capture–recapture studies. *Biometrics* **64** 377–385, 664. MR2432407

BRAVINGTON, M. (2015). Close-kin mark-recapture: Estimating the abundance of bluefin tuna from parent–offspring pairs. Preprint.

BRAVINGTON, M., GREWE, P. and DAVIES, C. (2014). Fishery-independent estimate of spawning biomass of Southern Bluefin Tuna through identification of close-kin using genetic markers. FRDC Report 2007/034, CSIRO, Australia.

COLE, D. J., MORGAN, B. J. T., CATCHPOLE, E. A. and HUBBARD, B. A. (2012). Parameter redundancy in mark-recovery models. *Biom. J.* **54** 507–523. MR2948080

CORMACK, R. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51** 429–438.

COTTERMAN, C. W. (1940). A calculus for statistico-genetics. Ph.D. thesis, Ohio State Univ., Columbus, OH.

CROW, J. F. and KIMURA, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, London. MR0274068

DAVEY, J. W., HOHENLOHE, P. A., ETTER, P. D., BOONE, J. Q., CATCHEN, J. M. and BLAXTER, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12** 499–510.

GRIEWANK, A. and WALTHER, A. (2008). *Evaluating Derivatives*: *Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, PA.

HILLARY, R. M., PREECE, A. L., DAVIES, C. R., KUROTA, H., SAKAI, O., ITOH, T., PARMA, A. M., BUTTERWORTH, D. S., IANELLI, J. and BRANCH, T. A. (2015). A scientific alternative to moratoria for rebuilding depleted international tuna stocks. *Fish and Fisheries*. DOI:10.1111/faf.12121.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

JARMAN, S. N., POLANOWSKI, A. M., FAUX, C. E., ROBBINS, J., PAOLI-ISEPPI, D., BRAVINGTON, M., DEAGLE, B. E. et al. (2015). Molecular biomarkers for chronological age in animal ecology. *Mol. Ecol.* **24** 4826–4847.

JOLLY, G. M. (1965). Explicit estimates from capture–recapture data with both death and immigration-stochastic model. *Biometrika* **52** 225–247. MR0210227

KANNO, Y., VOKOUN, J. C. and LETCHER, B. H. (2011). Sibship reconstruction for inferring mating systems, dispersal and effective population size in headwater brook trout (*Salvelinus fontinalis*) populations. *Conservation Genetics* **12** 619–628.

LANDER, E. S. and GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84** 2363–2367.

LUKACS, P. M. and BURNHAM, K. P. (2005). Review of capture–recapture methods applicable to noninvasive genetic sampling. *Mol. Ecol.* **14** 3909–3919.

MARSHALL, T. C., SLATE, J., KRUUK, L. E. B. and PEMBERTON, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7** 639–655.

MEAGHER, T. R. and THOMPSON, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Popul. Biol.* **29** 87–106. MR0830657

PALSBØLL, P., PEERY, Z. and BERUBE, M. (2010). Detecting populations in the 'ambiguous' zone: Kinship-based estimation of population structure at low genetic divergence. *Molecular Ecology Resources* **10** 797–805.

PEEL, D., BRAVINGTON, M. V., KELLY, N., WOOD, S. N. and KNUCKEY, I. (2013). A model-based approach to designing a fishery-independent survey. *J. Agric. Biol. Environ. Stat.* **18** 1–21. MR3067324

PLEDGER, S., POLLOCK, K. H. and NORRIS, J. L. (2010). Open capture–recapture models with heterogeneity: II. Jolly–Seber model. *Biometrics* **66** 883–890. MR2758224

POLANOWSKI, A. M., ROBBINS, J., CHANDLER, D. and JARMAN, S. N. (2014). Epigenetic estimation of age in humpback whales. *Molecular Ecology Resources*. **14** 976–987. DOI:10.1111/1755-0998.12247.

POLLOCK, K. H., NICHOLS, J. D., BROWNIE, C. and HINES, J. E. (1990). Statistical inference for capture–recapture experiments. *Wildlife Monographs* **107** 3–97.

SEBER, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika* **52** 249–259. MR0210228

SHENDURE, J. and JI, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* **26** 1135–1145.

SKARE, Ø., SHEEHAN, N. and EGELAND, T. (2009). Identification of distant family relationships. *Bioinformatics* **25** 2376–2382.

SKAUG, H. J. (2001). Allele-sharing methods for estimation of population size. *Biometrics* **57** 750–756. MR1859812

SKAUG, H., BERUBE, M. and PALSBØLL, P. (2010). Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. *Molecular Ecology Resources* **10** 693–700.

SOBEL, E., PAPP, J. C. and LANGE, K. (2002). Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70** 496–508.

THOMPSON, E. A. (1975). The estimation of pairwise relationships. *Ann. Hum. Genet.* **39** 173–188. MR0406572

THOMPSON, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics* **194** 301–326.

WANG, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics* **166** 1963–1979.

WANG, J. (2009). A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Mol. Ecol.* **18** 2148–2164.

WANG, J. (2010). Effects of genotyping errors on parentage exclusion analysis. *Mol. Ecol.* **19** 5061–5078.

WEIR, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

WRIGHT, J. A., BARKER, R. J., SCHOFIELD, M. R., FRANTZ, A. C., BYROM, A. E. and GLEESON, D. M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *Biometrics* **65** 833–840. MR2649856