

SPECIAL FEATURE

Ecological Perspectives of Pedigree Reconstruction with Genome-Wide Data

Comparison of kinship-identification methods for robust stock assessment using close-kin mark–recapture data for Pacific bluefin tuna

Yohei Tsukahara¹  | Reiichiro Nakamichi¹ | Aiko Matsuura¹ |
Tetsuya Akita¹  | Atushi Fujiwara² | Nobuaki Suzuki¹

¹Fisheries Resources Institute, Japan
Fisheries Research and Education Agency,
Yokohama, Kanagawa, Japan

²Fisheries Technology Institute, Japan
Fisheries Research and Education Agency,
Watarai, Mie, Japan

Correspondence

Yohei Tsukahara, Fisheries Resources
Institute, Japan Fisheries Research and
Education Agency, 2-12-4 Fukuura,
Kanazawa, Yokohama, Kanagawa
236-8648, Japan.
Email: tsukahara_yohei35@fra.go.jp

Funding information

JSPS KAKENHI, Grant/Award Numbers:
20KK0163, 23K05945; International
Resources Survey of the Fisheries Agency
of Japan

Abstract

Several attempts have been made to understand the population dynamics of fishery resources, such as tuna species using an integrated analysis model with multiple data sources. However, estimating absolute abundance levels in practical stock assessments remains a challenge. Close-kin mark–recapture (CKMR) methods provide information about the number of adults in a population using close-kinship pairs identified by genetic markers and statistical methods. In this study, we compared three methods for kinship identification using different algorithms in samples of wild Pacific bluefin tuna genotyped across 5029 genome-wide single nucleotide polymorphisms in 4108 samples. The flexible relationship analyzer by random forest method we developed employs pairwise identity-by-descent values as inputs for random forest classification. The other two methods were CKMRsim and COLONY, which have been published and applied in several studies. These three methods were applied to the actual genotyping data with moderate missing genotypes, in addition to the pseudo-generated genotyping data for the simulation test. The simulation test mimicked genotyping data with physical linkages as well as genetic characteristics similar to those of actual samples. The three methods resulted in different numbers of inferred kinship pairs for both generated and actual data. Particularly for the half-sibling pairs, a considerable number of false-positives and false-negatives existed in the identification results. The differences in kinship identification results were interpreted based on a simulation test. This study may enhance the understanding of how each software performs when applied to single-nucleotide polymorphism data with moderate missing genotypes, as demonstrated in this study.

KEYWORDS

close-kin mark–recapture, Pacific bluefin tuna, parental analysis, sibship reconstruction, stock size estimation

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Population Ecology* published by John Wiley & Sons Australia, Ltd on behalf of The Ecological Society of Japan.

1 | INTRODUCTION

Understanding population dynamics is the central focus of population ecology. Particularly, the vulnerability of wild species to exploitation, such as in fisheries, must be appropriately evaluated to conserve stocks while utilizing them effectively. The “integrated-analysis model” is often used as a fishery stock-assessment model; it employs the joint log-likelihood of several data sources, such as catch amounts, relative abundance indices, and size composition in the catch (Dichmont et al., 2016; Maunder & Punt, 2013). Consistent and accurate information on catch removal and the trends in stock abundance enable precise estimation of the absolute biomass level. However, most wild population estimates contain errors due to sampling limitation and model misspecification. Consequently, estimating absolute biomass in a fishery stock assessment model remains challenging (Deroba et al., 2014; Maunder & Piner 2015; Punt et al., 2018).

Recently, the close-kin mark-recapture (CKMR) method has attracted considerable attention (Bravington, Skaug, & Anderson, 2016; Skaug, 2017; reviewed in Bravington & Carroll, 2023; Casas & Saborido-Rey, 2023; Waples & Feutry, 2021). It was first used to assess marine fish stocks of southern bluefin tuna (SBT) and estimate the abundance of spawning stock biomass (SSB) (Bravington, Grewe, & Davies, 2016). Similar methods have been applied to cetacean species since the beginning of the 21st century (Nielsen et al., 2001; Skaug, 2001). The CKMR method commonly focuses on kinship pairs, including parent–offspring pairs (POPs), full-sibling pairs (FSPs), and half-sibling pairs (HSPs). This is analogous to the recapture of a marked individual using the traditional mark-recapture method.

The probability of finding close-kin pairs in a wild population can be employed to construct datasets for fishery stock assessment models and create likelihood functions to estimate the number of parents in the SSB using POPs or the number of reproductively active individuals of the population using HSPs (Bradford et al., 2018). We introduce a simple equation that does not consider the heterogeneity of fecundity among parents, such as age, body size, or residence time in the spawning grounds. In a wild population, the number of POPs found within a sample (k_{POP}) follows the binomial function:

$$k_{\text{POP}} \sim \text{Binom} \left[n_{\text{P}} \times n_{\text{O}}, \frac{2}{N} \right], \quad (1)$$

where n_{P} and n_{O} are the sample sizes of the parents and offspring, respectively, and N is the number of adult males and females in the population. The expected probability $2/N$ is a trivial case of sexual reproduction because

every child must have exactly two parents in the population. This formula provides an estimator of N in terms of $2n_{\text{P}}n_{\text{O}}/k_{\text{POP}}$ and yields the likelihood function (Bravington, Skaug, & Anderson, 2016). Fitting k_{POP} , which was conditioned by N (number of adults), to the observed distribution provides information for estimating the absolute scale of the integrated analysis.

Several studies have used CKMR to estimate the number of adults in wild populations. Marcy-Quay et al. (2020) and Prystupa et al. (2021) identified POPs in freshwater populations (brook trout and Arctic grayling) using microsatellites as molecular markers. These studies assumed a constant contribution to reproduction across parental individuals, and consequently, the probability of a sampled juvenile becoming a parent is homogeneous, regardless of individual fertility. Conversely, some studies have incorporated age- or body-size-specific fecundity into the POP probability. This incorporation reduces the likelihood of bias due to differences in reproductive contribution by age or body length (Bravington, Grewe, & Davies, 2016; Ruzzante et al., 2019; Sévêque et al., 2024; Wacker et al., 2021). However, considering age-specific fecundity typically requires precise age estimates for adult samples or the inclusion of the uncertainty of age estimates in the CKMR method (Petersma et al., 2024; Trenkel et al., 2022). Converting age from length measurement data with an age-length key is an alternative and simple way to deal with this model. However, an age-length key may introduce bias in age estimation, especially for species with relatively long life histories and flat-top growth curves. The use of the age-structured CKMR method is desirable when applying the CKMR method to wild species that have multiple spawning cycles during their lifetime (i.e., iteroparous species) and can be sampled over multiple cycles.

Additionally, introducing HSPs is useful not only for estimating the number of adults but also for directly determining biological parameters, such as mortality rates (Bravington, Skaug, & Anderson, 2016). This is because the probability of HSPs across different cohorts depends on the turnover of parents due to mortality and recruitment into the parental population. However, compared with POPs, identifying HSPs in a wild population is more difficult, requiring a larger set of markers to detect a smaller proportion of alleles identical by descent, and is complicated by the existence of other kin categories with similar degrees of genetic relatedness (Delaval et al., 2022; Hillary et al., 2018; McDowell et al., 2022; Patterson et al., 2022).

With the growing availability of next-generation sequencing, genome-wide single-nucleotide polymorphisms (SNPs) have become molecular markers that can be easily identified. Obtaining numerous SNP markers

enhances the ability to statistically identify kin relationships, even in the complex kin structure of a wild population; however, the larger the number of SNPs, the more physical linkages among them. Physically linked SNPs are often inherited together on the same chromosome, and this nonindependence means that the kinship information from each linked SNPs is reduced relative to that from independently segregating SNPs, which degrades the ability to identify kinship. Based on the number of molecular markers available, several methods use kinship indicators, such as pairwise relatedness coefficients (e.g., Huisman, 2017; Wang, 2022), pseudo-log-odds (Hillary et al., 2018; Lloyd-Jones et al., 2023), or reconstructed pedigree structures (Wang & Santure, 2009). Recently, Delaval et al. (2022) reported differences in the estimated number of HSPs using three identification methods. They used the HSPs identified by one software program, CKMRsim (Anderson, 2022), for downstream analysis because its identification results were the most conservative among the three methods. This indicates that no clear guidelines exist for selecting a method for kinship identification in CKMR studies of a specific target species.

In this study, we describe the kinship identification results obtained using three different methods: CKMRsim, the likelihood-ratio method applied in several CKMR studies (Delaval et al., 2022; Prystupa et al., 2021); COLONY (Jones & Wang, 2010), a family-based likelihood method, and flexible relationship analyzer by random forest (fraRF) (Nakamichi, 2024), a random forest classification method that uses identity-by-descent (IBD) information. We compared the methods using actual SNP data obtained from Pacific bluefin tuna (PBF; *Thunnus orientalis*) and simulated data generated to mimic the genomic characteristics of the actual PBF SNP data. This simulation test promotes an understanding of the advantages and disadvantages of each software application when applied to actual SNP data.

2 | MATERIALS AND METHODS

2.1 | PBF samples

The PBF is a highly migratory species, consisting of a single stock with two or more spawning grounds (Nakatsuka, 2020). It is a long-lived species with a lifespan of over 20 years, and individuals older than age 3 years can mature (Ashida et al., 2021; Bayliff, 1994). Two major spawning grounds for PBF are as follows: one in the waters around the Nansei Archipelago, hereafter called “Nansei,” and the other on the west side of the Sea of Japan, hereafter called “the SOJ.” According to

histological studies of PBF gonads, the major spawning season extends from April to July in Nansei (Ashida et al., 2015) and from July to August in the SOJ (Okochi et al., 2016). Longline fisheries in Nansei and purse-seine fisheries in the SOJ catch adult PBFs during the spawning season. Troll fisheries mainly catch age-zero PBFs for farming pens in nursery areas located around the spawning grounds. Although multiple spawning grounds generally make it difficult to apply the CKMR method, these fisheries enable the collection of putative parents and offspring with spawning ground information.

Muscle tissue samples for kinship identification were collected from harvests of each fishery in 2016. They were preserved in 6M TNES-UREA buffers (Asahida et al., 1996) immediately after sampling at the landing ports. The time duration from fish death to preservation depends on the characteristics of the fishery and the operational strategy, and because sampling was conducted voluntarily, some samples had relatively low-quality DNA in terms of fragmentation and concentration. We analyzed more than 4000 samples from the Nansei spawning grounds, in addition to a few samples from the SOJ. We extracted genomic DNA using the Maxwell® RSC Blood DNA kit (Promega Corporation, Madison, USA) or the DNeasy Blood and Tissue DNA extraction kit (Qiagen, Hilden, Germany) following the manufacturers' instructions. DNA samples were quantified using spectrophotometry (NanoDrop, Thermo Fisher Scientific, Waltham, MA, USA).

In total, we sequenced muscle tissues from 2437 juveniles and 1509 adults from the Nansei spawning grounds and 199 juveniles from the SOJ spawning grounds using GRAS-Di analysis and obtained robust genome-wide information from low-quality samples using random PCR-based genotyping (Enoki & Takeuchi, 2018). We outsourced the library preparation and next-generation sequencing (NGS) for GRAS-Di to Eurofins Genomics Inc. (Tokyo, Japan) and GeneBay Inc. (Kanagawa, Japan). In both cases, NGS was performed on a NovaSeq 6000 platform in 150 base paired-end mode (NovaSeq 6000 S4 Reagent Kit).

2.2 | Genotyping and variant calling

GRAS-Di sequencing data were trimmed and filtered using Trimmomatic ver 0.39 (Bolger et al., 2014) with the following parameters: ILLUMINACLIP and NexteraPE-PE. fa:2:30:10; HEADCROP: 5; SLIDING WINDOW: 4:20; MINLEN: 50. We used fastp ver 0.23.1 (Chen et al., 2018) with the following parameters: adapter_fasta NexteraPE-PE. fa, --average_qual 20, -q 20, -l 50, -t 5, and -T 5. Cleaned data were mapped to a reference genome

assembly (Accession No. DDBJ: BKCK01000001-BKCK01000444) using BWA-mem2 ver 2.2.1 (Vasimuddin et al., 2019) with default parameters. SAMtools (Danecek et al., 2021) were used to sort and index files from the binary alignment map before variant calling. We performed variant calling using the Genome Analysis Toolkit (GATK) HaplotypeCaller (McKenna et al., 2010) in the Genomic Variant Call Format (GVCF) mode for each sample. We then transferred the GVCF files of the 4145 samples to the GenomicsDB workspace using GenomicsDBImport and GenotypeGVCFs for joint genotyping.

We filtered the resulting VCF data using the following criteria: biallelic SNPs, $QD < 2.0$, $MQ < 40$, $MQRankSum < -12.5$, $ReadPosRankSum < -8.0$, $SOR > 4.0$, and $FS > 60.0$. After removing some samples suspected of artificial duplication/contamination or those with few heterozygous alleles (total number of heterozygous alleles < minor homozygous alleles), we filtered the VCF data further using VCFtools (Danecek et al., 2011) by setting the following parameters: `--max-alleles 2`, `--min-alleles 2`, `--minGQ 50`, `--minDP 20`, `--hwe 0.05`, `--max-missing 0.7`, `--maf 0.001`, `--recode`, `--remove-filtered-all`, `--non-ref-ac-any 1`, `--remove-indels`, `--exclude SNPs.txt` (which listed SNPs with observed heterozygosity > 0.5).

2.3 | Kinship-identification method

Two of the three methods used in this study have been published previously. One is “CKMRsim,” which uses likelihood-ratio methods to infer pairwise kinships and has been published on GitHub (<https://github.com/eriqande/CKMRsim/>). The likelihood method is a traditional method of kinship identification that sets a threshold. One of the features of CKMRsim is that the threshold can be set at an arbitrary point between two likelihood ratios to control the false-positive and false-negative rates. The control of the false-positive and false-negative rates contributes to reducing bias in the final CKMR estimates of demographic parameters (see details in the “Discussion” section). In this study, the thresholds for identifications of interested kinships from unrelated pairs were set at a false-positive rate of < 0.0001 to minimize false-positive identification as much as possible.

The likelihood ratio method determines kinship based on the fraction of the likelihood of two kinship categories, such as whether the pair is POP or an unrelated pair. This implies that CKMRsim must be applied multiple times to the dataset to identify the three types of kinship. As the comparison of log-likelihoods between

kinships of interest, especially for POPs from FSPs, can lead to more precise identification (Thompson & Meagher, 1987), we applied the sequential procedure in the following order. First, all possible pairs were tested to identify whether they were unrelated to or of interest to kinship, that is, POP, FSP, and HSP. In this procedure, a pair can be identified as having two or more specific kinships in duplicate. For example, a pair could be identified as both a POP and an FSP. Therefore, we applied subsequent identification for such pairs using a fraction of the kinships of interest; for example, the log-likelihood ratio between POP and HSP. In the comparison between two interesting kinships, the threshold was set as a point of intersection as no clear knowledge of the relative difference in appearance frequency between them was available.

As the log-likelihood ratio is a summation across markers, missing genotypes affect kinship identification results. While the training data used to determine the threshold were generated according to the maximum number of SNPs in the dataset, the number of SNPs for each pair varied according to the quality of the DNA samples. This could be the cause of misidentification. Therefore, we calculated some thresholds with fewer markers, and the log-likelihood ratio at those thresholds was linearly interpolated to predict the log-likelihood ratio values of the thresholds in accordance with the number of SNPs used in each pair (see details in Appendix S1). This method was proposed by the main developer of CKMRsim through personal communication to handle SNP data with numerous missing genotypes.

Another method was “COLONY (v2.0.6.6)” which is a family-based method that reconstructs the pedigrees in a sample using genetic data to determine the number of kinship pairs (Jones & Wang, 2010). The samples caught at each spawning ground were assumed to be of the parent generation, that is, both putative father and mother, and all samples, including both adult and juvenile samples, were assumed to be possible offspring. Given the characteristics of the dataset, which consisted of more than 4000 samples and genome-wide markers, the pairwise likelihood method in COLONY was applied for this analysis (Wang, 2012). This method required less time to reconstruct families compared with other time-consuming methods that may take a couple of weeks or even months.

In the case of Bluefin tuna, the expected family size (i.e., sibship size in COLONY) seemed small because of the relatively high amount of biomass. Therefore, we set the sibship size parameter to one under the constraint of a strong prior. The remaining parameters were set to default settings. Pairs with a probability of 100% were considered to be target kinships in this study. In addition,

we conducted 100 runs using different random seeds, which controlled the initial conditions for the estimation in the COLONY setting file, to find common pairs across the runs. While it has been reported that COLONY can lead to an increasing number of false-positive siblings as the sample size increases (Almudevar & Anderson, 2011), this procedure is recommended in the manual of this software to reduce false-positive misidentification, especially for HSP (see details in Appendix S2).

The third method was developed using a different approach. This method employs IBD relatedness as an input variable for a random forest algorithm called the fraRF to classify the pairs (Breiman, 2001; Nakamichi, 2024). Three scores for pairwise IBD relatedness (IBD0, IBD1, and IBD2) were calculated for each pair, using the maximum likelihood method (Milligan, 2003). IBD0 indicates the proportion of markers that shared neither allele identical by descent, whereas IBD1 and IBD2 are estimates of the proportions of markers that shared either one or both alleles identical by descent, respectively. An offspring inherits one of the alleles from one parent; thus, IBD1 in POPs is theoretically 1.0, whereas IBD0 and IBD2 are zero. As kinships with a substantial positive IBD2 score are rare, the explicit inclusion of IBD2 as the input for the random forest would enhance the ability of kinship identification, especially for distinguishing FSPs from other kinships.

The random forest function used in this analysis was in the R package “randomForest” (ver. 4.7–1.1). The number of trees and variables for each tree were 1000 and 2, respectively. No other parameters were changed in this function.

Although POPs and FSPs have relatively identifiable IBD scores, HSPs are difficult to identify because of their kinships with the same or similar IBD scores. For example, if either HSP has an offspring, then the offspring and the other HSP are half-sibling-derived uncle-nephews, which has a relatively similar IBD score to the HSPs. To distinguish distant kinships from HSPs, the random forest classification employed two distant kinship categories, that is, half-sibling-derived uncle/aunt-nephew/niece (hereafter referred to as half-thiatic pair [HTP]) and half-sibling-derived cousin (hereafter referred to as half-cousin pair [HCP]), in addition to the target kinships (see the next section for details).

2.4 | Data generation for training data

The two methods, CKMRsim and fraRF, require training data for identification, which involves the genetic characteristics of PBF. For CKMRsim, the software originally had a function to generate training data, namely, simulate_Qij. This illustrates the reference distribution of the

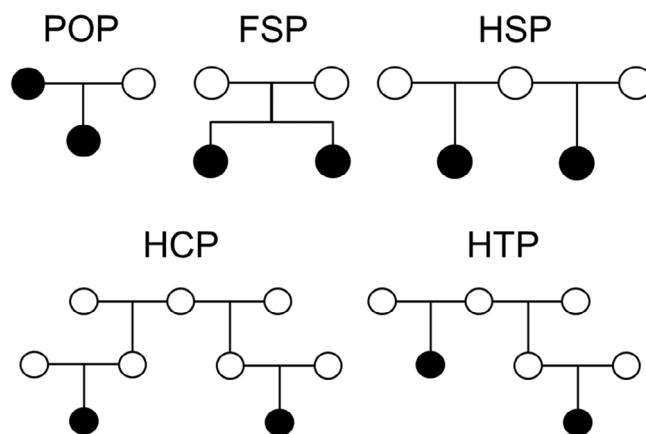


FIGURE 1 Schematic kin relationship of parent-offspring pair (POP), full-sibling pair (FSP), half-sibling pair (HSP), half-sibling derived cousin (HCP), and half-sibling derived thiatric pairs (HTP). Closed circles are in each kin relationship.

log-likelihood according to genetic characteristics in the input dataset. The threshold for identifying kinship was determined based on the reference distribution of training data. We generated 10,000 simulated samples as training data for CKMRsim. Conversely, fraRF employs an original data-generation protocol. The data generation for fraRF uses a simple family simulator consisting of two processes. The first process generated random genotypes of independent individuals based on the observed allele frequencies of the sample. The second process generated random haplotypes from the genotypes of the two individuals and combined them to generate the offspring genotype. In this generation, each locus was assumed to be independent, with no linkage between SNPs in the training data.

After generating the genotypes, we randomly selected 5% of the loci and replaced them with other random genotypes to mimic genotyping errors in the fraRF training data. Using these processes, we generated genotypes for POPs, FSPs, HSPs, and unrelated pairs (UP) in addition to two distant kinships (DKP), that is, HCP and HTP (Figure 1), according to the following definitions:

POP: Two independent parents were generated and mated to generate one offspring. One offspring and one of their parents were designated as a POP.

FSP: Two independent parents were generated and mated to generate two offspring. The POPs was designated as a FSP.

HSP: Three independent parents, A, B, and C, were generated, and parent A mates with parents B and C to generate one offspring each. This pair of offspring was designated as the HSP.

HCP: Three independent parents, A, B, and C, were generated. Parent A mates with parents B and C to

generate offspring X and Y, respectively, which means that X and Y are HSP. Two additional independent parents, D and E, were generated, and offspring X mated with parent D and offspring Y mated with parent E to generate one grand offspring each. The pair of grand offspring, that is, half-sibling-derived cousins, is designated as the HCP.

HTP: Three independent parents, A, B, and C, were generated. Parent A is mated with parent B to generate offspring X, and parent A is mated with parent C to generate offspring Y, meaning that X and Y are HSP. An additional independent parent, D, mates with X to generate a grand offspring. The pair between the grand offspring and Y, that is, half-sibling-derived thiatric (uncle-nephew, aunt-niece, etc.), is designated as the HTP.

UP: Two independent individuals were generated and designated as the UP.

Independent simulations were performed for each of the following six kinship pairs: POP, FSP, HSP, DKP (HCP and HTP), and UP. We generated 500 pairs for each kinship type to be used as training data for fraRF with a 1% genotyping error. The number of generated pairs can be set arbitrarily.

2.5 | Data generation for simulation test

We conducted a simulation test to facilitate the interpretation of the identification results across the three software packages. The input dataset for the simulation test was generated in a two-step manner for the training data in fraRF, whereas the generated SNP data had genetic characteristics similar to those of the actual PBF genotyping data in terms of the missing rate, minor allele frequency (MAF), and physical linkages. To account for the physical linkage in SNP generation, the origin genotypes in the first process of SNP generation contained information on the chromosomes.

Cytological studies have reported 24 pairs of chromosomes in *Thunnus* species (Ida et al., 1991; Soares et al., 2013). Based on the number of chromosomes, a sex-specific genetic map of PBF with 24 linkage groups was constructed using the draft genome (Uchino et al., 2016, 2018). As a result, the average size of linkage groups was 96.6 centimorgan (cM) for males and 105.7 cM for females, indicating that the recombination would occur once at each chromosome across one generation. We assumed that the SNPs were uniformly distributed on the 48 chromosomes. During gamete formation, each pair of chromosomes generates four haplotypes. Recombination occurred in two of the four haplotypes at a certain site. The position of the recombination site in

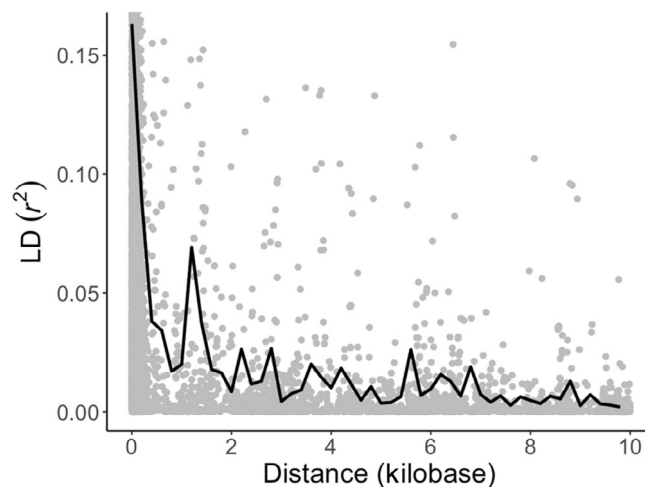


FIGURE 2 Linkage equilibrium decay plot for the dataset of Pacific bluefin tuna. Shaded gray points and black line indicate raw values and 200 base pair averaged values, respectively.

each haplotype was randomly determined using a uniform probability distribution. One of the four haplotypes is randomly inherited from a parent (either the father or mother) to the next generation. This process, which incorporates chromosomal information, can simulate the physical linkages of SNPs.

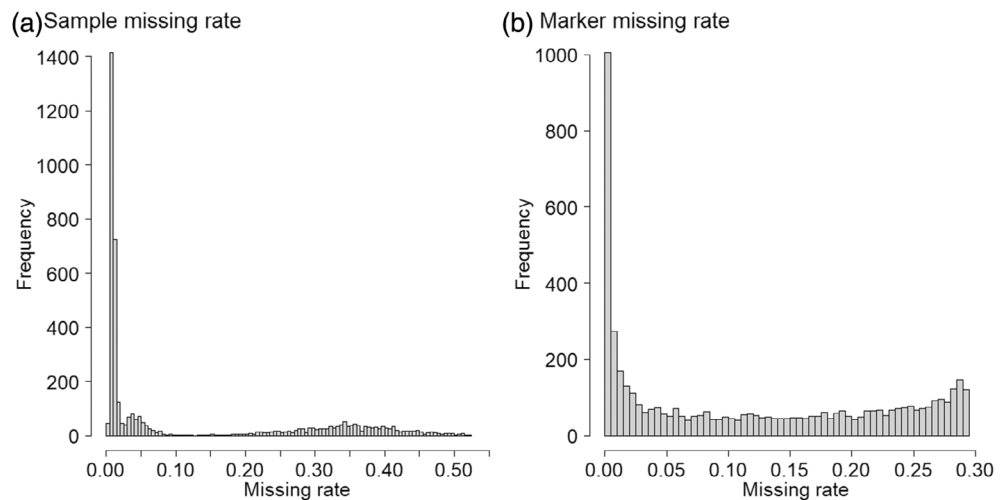
The allele frequency in the original genotypes and missing rates in genotyping across samples and markers were adjusted to be as similar as possible to the observed values. In addition, the existence rates of specific kinships (POPs, FSPs, and HSPs) were assumed to be the same as the average probability identified in the actual PBF identification across the three software packages.

3 | RESULTS

3.1 | Genotyping results

Genotyping using GRAS-Di and data filtering resulted in 62,946 polymorphic SNPs in 4111 samples with missing genotyping data. Some of the SNPs in such large numbers were physically linked to each other, which reduced the accumulation of statistical power for kinship identification with each additional SNP. The relationship between genetic distance and linkage disequilibrium (LD) between the remaining SNPs was investigated (Figure 2). As a result, SNPs with LD, that is, those with $r^2 > 0.01$, and those with a distance of less than 5000 base pairs (bp) were filtered out using PLINK v. 1.90 (Purcell et al., 2007) with the following parameters: --indep-pairwise 1000 100 0.01 and --thin 5000. In addition, SNPs with MAF of less than 0.05 were excluded to avoid

FIGURE 3 Histogram of the missing rate of genotyping in 5029 SNPs across markers (left) and samples (right) in 4108 samples of Pacific bluefin tuna.



genotyping errors in the dataset. Finally, three samples were removed from this analysis because they showed exceptionally high heterozygosity, with F -values (the inbreeding coefficient) less than -0.5 , which were calculated using PLINK. The filtered genotyped data contained 5029 SNPs from 4108 samples remaining for kinship identification.

The histogram of missing genotype rates across the samples showed a trimodal distribution, with peaks at approximately 0%, 5%, and 35% (Figure 3). This indicated that more than half of the samples had good DNA quality with few missing genotypes, whereas the rest had fragmented DNA, resulting in moderate missing rates across the markers (Figure 3a). The histogram of missing rates across the markers also showed a skewed distribution, indicating that some loci were specifically lacking and that the missing loci were not randomly distributed in the obtained SNPs. Moderate and biased missing patterns affect the kinship identification results.

3.2 | Kinship identification in PBF

The COLONY method identified 23 POPs, whereas fraRF and CKMRsim identified 26 and 25 POPs, respectively (Table 1). The multiple runs using different random seeds in COLONY lost one pair in POP identification. The 23 pairs were identically classified as POPs across the three different models and no unique pair was present in COLONY identification, whereas one pair was uniquely identified as POPs by fraRF, and two pairs were identified as POPs by fraRF and CKMRsim (Figure 4). One of two pairs identified as POPs by fraRF and CKMRsim was dropped by COLONY over the course of multiple runs. Interestingly, offspring in these two candidate POPs was the same, suggesting that both the mother and father were caught in different years and different

TABLE 1 Number of parent–offspring pairs (POPs), full-sibling pairs (FSPs), and half-sibling pairs (HSPs) inferred by the three different algorithms in the Pacific bluefin tuna samples.

Kinship	CKMRsim	COLONY	fraRF
POP	25	23	26
FSP	13	14	14
HSP	296	351	324

ports. The lack of sex information of parent samples and the assumption that parent samples are both father and mother candidates might disturb the identification of these pairs in COLONY. In addition, one pair identified as POP by only fraRF was identified as HSP by the other two methods. The IBD_0 value in this pair was approximately 0.30, which was likely too high to be POP, suggesting fraRF may be wrong.

There were slight differences in the identification results for the FSPs: 13 for CKMRsim and 14 for fraRF and COLONY (Table 1). FSP identification by different random seeds in COLONY was consistent across 100 runs. Thirteen of the 14 FSPs were consistently identified by all software and were pairs of age-zero fish born in the same year. The other FSP in COLONY and fraRF was identified as HSPs by CKMRsim, which was also a pair of age-zero fish born in the same year (Figure 4). The IBD_2 value of this pair was 0.13. This was a substantial value for IBD_2 , although it was lower than the theoretical value of the FSP. Identification of the FSP showed relatively comparable performance across the software probably due to the unique genetic characteristics.

Furthermore, the differences in the number of inferred HSPs were larger than those in POPs and FSPs (Table 1). CKMRsim identified 296 HSPs, whereas COLONY and fraRF identified 351 and 324 HSPs, respectively. In terms of the multiple runs using different

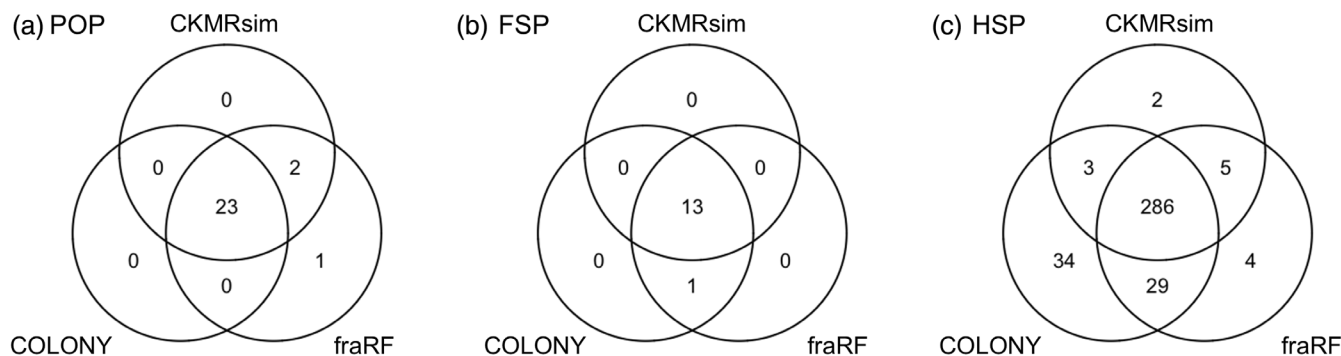


FIGURE 4 Venn diagram of the inferred number of (a) parent-offspring pair (POP), (b) full-sibling pair (FSP), and (c) half-sibling pair (HSP) by three software.

random seeds in COLONY, no runs inferred less than 390 HSPs, supporting the idea that COLONY with a single run tends to overestimate sibships. The number of common pairs across different random seed settings gradually decreased naturally, and finally, 351 HSPs remained after 100 runs. However, the number of HSPs in COLONY was still the highest among the three software. The number of common pairs could further decrease by other random seeds (see detail in Appendix S2). Significant differences in HSP identification can lead to considerable differences in the estimated adult population sizes among the three methods. CKMRsim produced the lowest values for both FSPs and HSPs, whereas kinship identification with the likelihood ratio method often employs additional adjustment of the number of HSPs using the false-negative rate (Bravington, Skaug, & Anderson, 2016) (see details in the “Discussion” section).

3.3 | Kinship identification in simulation test

Based on the genotyping and kinship identification results for the PBF samples, a pseudo-genotyping dataset was generated using 2000 samples and 6000 SNPs for the simulation test. The MAF for the generated data was assumed to be the same as the SNPs used for the actual PBF identification. As the missing rates across the samples and markers were skewed, we employed a three-step procedure to mimic those in the actual sample. First, half of the markers had a uniform missing rate of 0%–1%, and the remaining markers had a uniform missing rate of 1%–30%. Half of the samples had low probabilities of missing genotyping data, 20% had moderate probabilities, and the remainder had high probabilities. Finally, the missing genotypes at each locus were assigned to each sample according to their weighted values (Figure 5). The number of kinship pairs was assumed to

be close to the average values across the three methods: six for POPs, three for FSPs, and 80 for HSPs. We assumed that the samples contained the same number of HCPs and HTPs as HSPs.

A simulation test using the generated genotypes showed that fraRF and COLONY correctly identified all POPs and FSPs (Table 2), whereas CKMRsim had one false-negative misidentification of both POPs and FSPs. However, every software program misidentified the HSPs for both false-negative and false-positive results, indicating the difficulty of HSP identification, even with numerous SNPs. Sixty-eight pairs were correctly identified as HSP by CKMRsim, 72 using COLONY, and 74 using fraRF. The number of pairs identified as HSP, including misidentifications, was 74, 79, and 82 for CKMRsim, COLONY, and fraRF, respectively. No software could perform HSP identification perfectly.

4 | DISCUSSION

For fishery stocks, sampling generally relies on harvests from commercial fisheries. Collecting well-preserved samples from commercial fisheries is difficult; thus, sample quality tends to be moderate. There is a trade-off between sample size and quality because CKMR studies generally require a massive effort to find the target number of kinship pairs according to the size of the wild population, although the accuracy of kinship identification is highly dependent on sample quality. In this study on PBF, we prioritized the use of as many samples as possible, rather than selecting good-quality samples, to increase the number of candidate pairs. Owing to the GRAS-Di technique, we obtained many markers even from relatively low-quality samples, and there were many missing genotypes.

Including low-quality samples reduced the number of common markers used in each pairwise comparison. As

FIGURE 5 Histogram of the missing rate of genotyping in 6000 SNPs across markers (left) and samples (right) in 2000 samples of generated data for simulation test. SNPs, single-nucleotide polymorphisms.

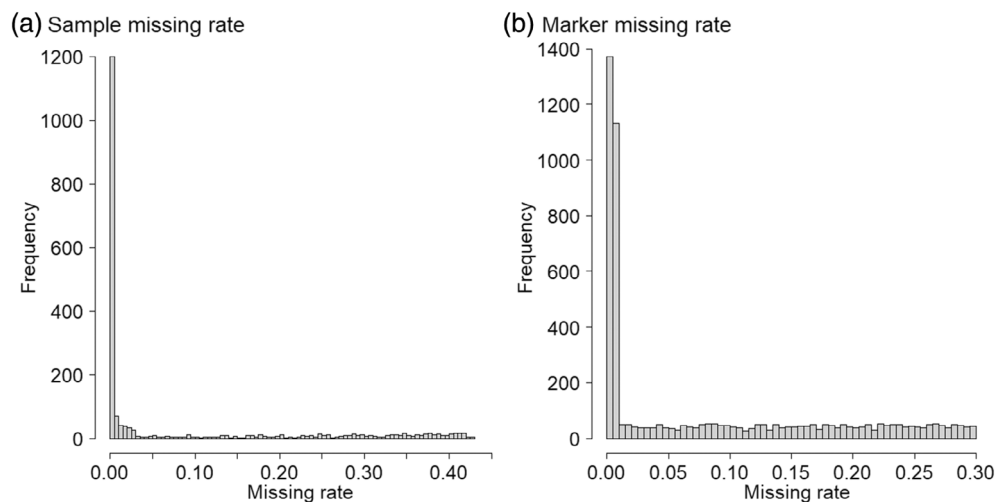


TABLE 2 Number of parent-offspring pairs (POPs), full-sibling pairs (FSPs), and half-sibling pairs (HSPs) inferred by the three different algorithms in the generated data for simulation test.

Kinship	CKMRsim	COLONY	fraRF	True number
POP	5 (0, -1)	6 (0, -0)	6 (0, -0)	6
FSP	2 (0, -1)	3 (0, -0)	3 (0, -0)	3
HSP	74 (6, -12)	79 (7, -8)	82 (8, -6)	80

Note: The positive number in the parenthesis indicates the number of false-positive identifications, and the negative one indicates the false-negative identifications.

COLONY uses the maximum likelihood method to reconstruct the family and fraRF uses it to calculate the IBD values, a smaller number of markers would make the inference uncertain, however, not biased. On the other hand, the log-likelihood ratio method in CKMRsim calculates the sum of the ratio over the used marker, and hence, the estimated values from different numbers of used markers must be biased. Therefore, we applied dynamic thresholds that varied linearly with the number of markers used for each pair. This enhanced the kinship identification ability of CKMRsim. However, multiple calculations were required to generate knots for linear interpolation. In this study, we conducted five or six calculations for each comparison, for example, POP versus UP and HSP versus HTP. The multiple calculations would lose one of the great advantages of the log-likelihood ratio method, that is, the time-saving method (see details in Appendix S1).

Concerning the trade-off between false-positive and false-negative rates, the false-positive rate should be sufficiently low for the application of CKMR to a large population as there are far larger numbers of non-target kinships (e.g., unrelated pairs) than target pairs (i.e., POPs and HSPs). Therefore, the number of false-positives is naturally much higher than that of false-negatives under the assumption of the same false-positive and false-negative rates. Setting the threshold for false-positive rate to 0.0001 for CKMRsim, which is the

minimum value for the default options, can be considered reasonable in this study with the PBF, which seems to have a large population. However, even with this strict threshold, there would be several false-positive misidentified pairs because 4108 samples made approximately 8.5 million pairs; thus, more than 850 false-positive misidentifications could theoretically occur. Although a lower false-positive rate might be desirable according to the number of samples, moderate false-negative misidentifications were observed for POPs and FSPs in the simulation test. False-negative misidentification is unavoidable to some extent.

Despite the complexity of the threshold issue, one of the advantages of defining it is that the theoretical false-negative rate can be theoretically conditioned and preliminarily estimated (Bravington, Skaug, & Anderson, 2016). Recently, CKMR modeling, which considers the false-negative rate, has been applied to demographic parameter estimates when it is assumed that there are no or few false-positive misidentifications (Patterson et al., 2022; Punt et al., 2024). Explicit consideration of the false-negative rates, for example, dividing the identified number by one minus the false-negative rate, would improve the accuracy of the identification. However, no parameters or settings were available to control for false-positive and false-negative rates for COLONY and fraRF. Only CKMRsim can control the false identification rates among the three software in this study.

By contrast, COLONY and fraRF showed slightly better identification results in the simulation test for all kinships of interest. At least, COLONY and fraRF identified POPs and FSPs correctly. As previously mentioned, COLONY tends to overestimate sibships when the population size is large. In this case, despite assuming a large population in the simulation test, the common pairs across 100 runs with different random seeds reduced false-positive identifications, whereas every run showed overestimation with false-positive misidentifications. While COLONY tended to overestimate the number of HSPs in each run, seeking common pairs across different random seed settings would reduce the degree of overestimation, even for large populations with sparse sampling. However, overlooking HSPs in any run can lead to false-negative misidentification. This method requires many runs to find a stable bottom-out and can be time-consuming.

fraRF showed equivalent performance in the simulation test compared with the other two software. The identified numbers of POPs and HSPs were exact, even with many missing genotyping data points. This method uses a set of pairwise IBD values calculated from the available SNPs of each pair and a maximum likelihood estimation, which converts raw genetic information into a statistical indicator. Owing to the high classification power of the random forest, fraRF with simply three IBD values showed comparable results with the other software in the simulation tests by only a one-time calculation. However, this method has overestimated HSPs so far, due to moderate misidentification of HTP and HSP (see details in Appendix S3). According to the PBF life history and maturity schedule, there should be HTP in the wild population.

Despite the current simple design of fraRF with IBD values and random forest, the simulation test revealed that the identification ability of fraRF was not much different from the other methods. This could be a candidate method to use the kinship identification for the CKMR study at least for the POP and FSP as well as the other software. To enhance the identification ability, particularly for HSP, fraRF could be modified by including some new input variables even if they are not genetic information. For example, loglikelihood ratio values and the number of markers used are candidates to improve the performance of this software.

In this study, we have generated 1000 individuals and 500 pairs for each kinship in training data. The number of training data was smaller than that for CKMRsim, while increasing the amount of training data can lead to having pairs with rare but not extreme IBD values, resulting in a more representative training dataset. The dataset for PBF in this study had a large number of SNPs for the IBD

values to be distinguishable enough. Also, the calculation of IBD values is computationally intensive. These were the reasons the number of training data was set to 500 in this study. However, if the number of available SNPs or other information for fraRF is limited, the number of training data should be raised accordingly for fraRF to avoid inaccuracies. Implementing an efficient method to reduce the calculation for training data, such as importance sampling (Anderson & Garza, 2006), might be useful to enhance the usability of the fraRF for other cases.

This simulation test did not include grandparent-grand offspring kinship (GG), which should disturb the identification of HSP due to the same IBD values as HSPs. The degradation of the identification performance for the actual PBF dataset is inevitable because of the existing GG pairs in the actual samples. This is a limitation of this software, and identification would require more information to distinguish HSPs from GG. External biological information, such as age gaps or sizes between individuals in pairs, can help identify them. Although additional simulation models are necessary to generate training data with combined genetic and biological information, including these nongenetic variables in the random forest may enhance the identification performance.

The identification results of the actual PBF samples exhibited a similar tendency to those of the simulation tests (Table 1). CKMRsim had smaller numbers of POP, FSP, and HSP, which may have been due to its strict threshold. For HSP, the three methods showed large discrepancies similar to the simulation results. CKMRsim underestimated the number of HSPs as intended, while it had some false-positive misidentifications. On the other hand, COLONY and fraRF accepted a somewhat balanced number of false-negative and false-positive misidentifications, resulting in the seemingly close to the true number. It should be noted that no software could get the correct number of HSPs, indicating the difficulty of identifying this relationship even with the high number of SNPs. The physical linkage among the SNPs reduced the identification power of every three software. Using the number of HSP in the CKMR estimation, the user should consider the uncertainty of the kinship identification and post hoc approach, for example, adjustment by false-negative rate, would enhance the accuracy of kinship identification.

CKMR can be used for many fishery stock assessments in the near future. Kinship identification is a key technique used in CKMR applications. However, this process is strongly affected by upstream analysis, such as genotyping, and impacts downstream analysis. In this study, the quality of the PBF data was generally adequate, although poor for DNA sequencing in some samples. As a result, the missing rate of genotypes made

kinship identification more difficult. In addition, no software correctly estimated the number of HSP in this simulation test using the mimicked physical linkage and missing rate. Based on the findings in this study, the kinship identification method should be appropriately used in accordance with the genetic characteristics of the target species for CKMR studies.

AUTHOR CONTRIBUTIONS

Yohei Tsukahara and Tetsuya Akita conceived the idea of this study. Yohei Tsukahara drafted the original manuscript. Atushi Fujiwara conducted the creation of input data for analysis. Reiichiro Nakamichi and Aiko Matsuura developed and conducted a kinship-identification analysis. Yohei Tsukahara, Tetsuya Akita, and Reiichiro Nakamichi contributed to the interpretation of the results. Nobuaki Suzuki supervised the conduct of this study. All authors reviewed the manuscript draft, revised it critically, and approved the final version for submission.

ACKNOWLEDGMENTS

We would like to thank Dr. Shuya Nakatsuka and Dr. Hiromu Fukuda for their valuable comments. We also gratefully acknowledge all the people who were involved in this study to collect the tissue samples. In addition, we would like to thank Dr. Jing Wang (COLONY developer and handling editor) and Dr. Eric Anderson (CKMRsim developer and reviewer) for giving us many useful comments and suggestions to use their software. Also, we deeply thank Dr. Pierre Feutry for giving us insightful comments as one of the reviewers. Further, we appreciate efforts to incorporate our dataset into the Kinference by Dr. Baylis Shane and Dr. Mark Bravington, but it could not be accomplished due to our time constraints. This research was financially supported by the Promotion Program for the International Resources Survey of the Fisheries Agency of Japan and JSPS KAKENHI Grant Number 20KK0163 and 23K05945.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The fraRF software developed by R code and generated genotypes for the simulation test is available in the Zenodo (<https://zenodo.org/records/13383270>). The datasets of actual genotypes in this study are also available from the corresponding author upon reasonable request.

ORCID

Yohei Tsukahara  <https://orcid.org/0009-0001-5989-9466>

Tetsuya Akita  <https://orcid.org/0000-0003-4940-886X>

REFERENCES

- Almudevar, A., & Anderson, C. E. (2011). A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. *Molecular Ecology Resources*, 12(1), 164–178. <https://doi.org/10.1111/j.1755-0998.2011.03061.x>
- Anderson, E. C. (2022). *CKMRsim*. <https://github.com/eriqande/CKMRsim/>
- Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, 172(4), 2567–2582. <https://doi.org/10.1534/genetics.105.048074>
- Asahida, T., Kobayashi, T., Saitoh, K., & Nakayama, I. (1996). Tissue preservation and Total DNA extraction from fish stored at ambient temperature using buffers containing high concentration of Urea. *Fisheries Science*, 62, 727–730. <https://doi.org/10.2331/fishsci.62.727>
- Ashida, H., Okochi, Y., Oshimo, S., Sato, T., Ishihara, Y., Watanabe, S., Fujioka, K., Furukawa, S., Kuwahara, T., Hiraoka, Y., & Tanaka, Y. (2021). Differences in the reproductive traits of Pacific bluefin tuna *Thunnus orientalis* among three fishing grounds in the sea of Japan. *Marine Ecology Progress Series*, 662, 125–138.
- Ashida, H., Suzuki, N., Tanabe, T., Suzuki, N., & Aonuma, Y. (2015). Reproductive condition, batch fecundity, and spawning fraction of large Pacific bluefin tuna *Thunnus orientalis* landed at Ishigaki Island, Okinawa, Japan. *Environmental Biology of Fishes*, 98(4), 1173–1183. <https://doi.org/10.1007/s10641-014-0350-8>
- Bayliff, W. H. (1994). A review of the biology and fisheries for northern bluefin tuna, *Thunnus thynnus*, in the Pacific Ocean. FAO Fisheries Technical Paper 336/2, 244–295.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bradford, R. W., Thomson, R., Bravington, M., Foote, D., Gunasekera, R., Burece, B., Harasti, D., Otway, N., & Feutry, P. (2018). *A close-kin mark-recapture estimate of the population size and trend of east coast grey nurse shark*. Report to the National Environmental Science Program, Marine Biodiversity Hub. CSIRO Oceans & Atmosphere.
- Bravington, M. V., & Carroll, E. L. (2023). The practical magic of close-kin mark-recapture. In O. F. Berry, E. C. Holleley, & S. N. Jarman (Eds.), *Applied environmental genomics*. CISRO Publication, Chapter 11.
- Bravington, V. M., Grewe, M. P., & Davies, R. C. (2016). Absolute abundance of southern bluefin tuna estimated by close-kin mark-recapture. *Nature Communications*, 7, 13162. <https://doi.org/10.1038/ncomms13162>
- Bravington, V. M., Skaug, J. H., & Anderson, C. E. (2016). Close-kin mark-recapture. *Statistical Science*, 31, 259–274. <https://doi.org/10.1214/16-STS552>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Casas, L., & Saborido-Rey, F. (2023). A review of an emerging tool to estimate population parameters: The close-kin mark-recapture method. *Frontiers in Marine Science*, 10. <https://doi.org/10.3389/fmars.2023.1087027>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), 884–890. <https://doi.org/10.1093/bioinformatics/bty560>

- Danecek, P., Auton, A., Abecasis, G., Albers, A. C., Banks, E., DePristo, A. M., Handsaker, E. R., Lunter, G., Marth, T. G., Sherry, T. S., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Danecek, P., Bonfield, K. J., Liddle, J., Marshall, J., Ohan, V., Pollard, O. M., Whitwham, A., Keane, T., McCarthy, A. S., Davies, M. R., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.1093/gigascience/giab008>
- Delaval, A., Bendall, V., Hetherington, S. J., Skaug, H. J., Frost, M., Jones, C. S., & Noble, L. R. (2022). Evaluating the suitability of close-kin mark-recapture as a demographic modelling tool for a critically endangered elasmobranch population. *Evolutionary Applications*, 16, 1–13.
- Deroba, J. J., Butterworth, S. D., Methot, D. R., De Oliveira, A. A. J., Fernandez, C., Nielsen, A., Cadrin, X. S., Dickey-Collas, M., Legault, C. M., Ianelli, J., Valero, L. J., Needle, L. C., O'Malley, M. J., Chang, Y., Thompson, G. G., Canales, C., Swain, P. D., Miller, C. M. D., Hintzen, T. N., ... Hulson, J. F. P. (2014). Simulation testing the robustness of stock assessment models to error: Some results from the ICES strategic initiative on stock assessment methods. *ICES Journal of Marine Science*, 72, 19–30. <https://doi.org/10.1093/icesjms/fst237>
- Dichmont, M. C., Deng, A. R., Punt, E. A., Brodziak, J., Chang, Y., Cope, M. J., Ianelli, N. J., Legault, M. C., Methot, D. R., Porch, E. C., Prager, H. M., & Shertzer, W. K. (2016). A review of stock assessment packages in the United States. *Fisheries Research*, 183, 447–460. <https://doi.org/10.1016/j.fishres.2016.07.001>
- Enoki, H., & Takeuchi, Y. (2018). New genotyping technology, GRAS-Di, using next generation sequencer. *Proceedings of the plant and animal genome conference XXVI*. San Diego, CA. <https://pag.confex.com/pag/xxvi/meetingapp.cgi/Paper/29067>
- Hillary, R. M., Bravington, M. V., Patterson, T. A., Grewe, P., Bradford, R., Feutry, P., Gunasekera, R., Peddemors, V., Werry, J., Francis, M. P., Duffy, C. A. J., & Bruce, B. D. (2018). Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand. *Scientific Reports*, 8(1), 2661. <https://doi.org/10.1038/s41598-018-20593-w>
- Huisman, J. (2017). Pedigree reconstruction from SNP data: Parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, 17, 1009–1024. <https://doi.org/10.1111/1755-0998.12665>
- Ida, H., Oka, N., & Hayashigaki, K. (1991). Karyotypes and cellular DNA contents of three species of the subfamily Clupeinae. *Japanese Journal of Ichthyology*, 38, 289–294.
- Jones, O., & Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10, 551–555. <https://doi.org/10.1111/j.1755-0998.2009.02787>
- Lloyd-Jones, L. R., Bravington, M. V., Armstrong, K. N., Lawrence, E., Feutry, P., Todd, C. M., Dorrestein, A., Welbergen, J. A., Martin, J. M., Rose, K., Hall, J., Phalen, D. N., Peters, I., Baylis, S. M., Macgregor, N. A., & Westcott, D. A. (2023). Close-kin mark-recapture informs critically endangered terrestrial mammal status. *Scientific Reports*, 13, 12512. <https://doi.org/10.1038/s41598-023-38639-z>
- Marcy-Quay, B., Sethi, A. S., Therkildsen, O. N., & Kraft, E. C. (2020). Expanding the feasibility of fish and wildlife assessments with close-kin mark-recapture. *Ecosphere*, 11, e03259. <https://doi.org/10.1002/ecs2.3259>
- Maunder, N. M., & Piner, R. K. (2015). Contemporary fisheries stock assessment: Many issues still remain. *ICES Journal of Marine Science*, 72, 7–18. <https://doi.org/10.1093/icesjms/fsu015>
- Maunder, N. M., & Punt, E. A. (2013). A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142, 61–74. <https://doi.org/10.1016/j.fishres.2012.07.025>
- McDowell, J. R., Bravington, M., Grewe, P. M., Lauretta, M., Walter, J. F., III, Baylis, S. M., Gosselin, T., Malca, E., Gerard, T., Shiroza, A., Lamkin, J. T., Biesack, E. E., Zapfe, G., Ingram, W., Davies, C., & Porch, C. (2022). Low levels of sibship encourage use of larvae in western Atlantic bluefin tuna abundance estimation by close-kin mark-recapture. *Scientific Reports*, 12(1), 18606.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Daly, G. S., & DePristo, A. M. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, 163, 1153–1167.
- Nakamichi, R. (2024). *Supplementary information for: Comparison of kinship-identification methods for robust stock assessment using close-kin mark-recapture data for Pacific bluefin tuna*. <https://zenodo.org/records/13383270>
- Nakatsuka, S. (2020). Stock structure of Pacific bluefin tuna (*Thunnus orientalis*) for management purposes – A review of available information. *Reviews in Fisheries Science & Aquaculture*, 28, 170–181. <https://doi.org/10.1080/23308249.2019.1686455>
- Nielsen, R., Mattila, D. K., Clapham, P. J., & Palsbøll, P. J. (2001). Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics*, 157(4), 1673–1682.
- Okochi, Y., Abe, O., Tanaka, S., Ishihara, Y., & Shimizu, A. (2016). Reproductive biology of female Pacific bluefin tuna, *Thunnus orientalis*, in the sea of Japan. *Fisheries Research*, 174, 30–39. <https://doi.org/10.1016/j.fishres.2015.08.020>
- Patterson, T. A., Hillary, R. M., Kyne, P. M., Pillans, R. D., Gunasekera, R. M., Marthick, J. R., & Feutry, P. (2022). Rapid assessment of adult abundance and demographic connectivity from juvenile kin pairs in a critically endangered species. *Science Advances*, 8(51), eadd1679.
- Petersma, F. T., Thomas, L., Harris, D., Bradley, D., & Papastamatiou, Y. P. (2024). Age is not just a number: How incorrect ageing impacts close-kin mark-recapture estimates of population size. *Ecology and Evolution*, 14(6), e11352. <https://doi.org/10.1002/ece3.11352>
- Prystupa, S., McCracken, R. G., Perry, R., & Ruzzante, E. D. (2021). Population abundance in arctic grayling using genetics and close-kin mark-recapture. *Ecology and Evolution*, 11, 4763–4773. <https://doi.org/10.1002/ece3.7378>
- Punt, A. E., Day, D., Fay, G., Haddon, M., Klaer, N., Little, L. R., Johnson, K. P., Smith, D. M. A., Smith, D. C., Sporcic, M.,

- Thomson, R., Tuck, G. N., Upston, J., & Wayte, S. (2018). Retrospective investigation of assessment uncertainty for fish stocks off southeast Australia. *Fisheries Research*, 198, 117–128. <https://doi.org/10.1016/j.fishres.2017.10.007>
- Punt, A. E., Thomson, R., Little, L. R., Bessell-Browne, P., Burch, P., & Bravington, B. (2024). Including close-kin mark-recapture data in statistical catch-at-age assessments and management strategies. *Fisheries Research*, 276, 107057. <https://doi.org/10.1016/j.fishres.2024.107057>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Ruzzante, E. D., McCracken, R. G., Førlund, B., MacMillan, J., Notte, D., Buhariwalla, C., Flemming, M. J., & Skaug, H. (2019). Validation of close-kin mark-recapture (CKMR) methods for estimating population abundance. *Methods in Ecology and Evolution*, 10, 1445–1453. <https://doi.org/10.1111/2041-210X.13243>
- Sévêque, A., Robert, L., Lisette, W., Kristin, B., Lisa, K., Caitlin, O., Sarah, M., Norton, D., Petroelje, T., Swenson, D., & Morin, D. (2024). Sources of bias in applying close-kin mark-recapture to terrestrial game species with different life histories. *Ecology*, 105(3), e4244. <https://doi.org/10.1002/ecy.4244>
- Skaug, H. J. (2001). Allele-sharing methods for estimation of population size. *Biometrics*, 57(3), 750–756.
- Skaug, H. J. (2017). The parent-offspring probability when sampling age-structured populations. *Theoretical Population Biology*, 118, 20–26.
- Soares, R. X., Bertollo, L. A. C., Costa, G. W. W. F., & Molina, W. F. (2013). Karyotype stasis in four Atlantic Scombridae fishes: Mapping of classic and dual-color FISH markers on chromosomes. *Fisheries Science*, 79, 177–183. <https://doi.org/10.1007/s12562-013-0602-0>
- Thompson, E. A., & Meagher, T. R. (1987). Parental and sib likelihoods in genealogy reconstruction. *Biometrics*, 43(3), 585–600.
- Trenkel, V. M., Charrier, G., Lorange, P., & Bravington, M. V. (2022). Close-kin mark-recapture abundance estimation: Practical insights and lessons learned. *ICES Journal of Marine Science*, 79(2), 413–422. <https://doi.org/10.1093/icesjms/fsac002>
- Uchino, T., Hosoda, E., Nakamura, Y., Yasuike, M., Mekuchi, M., Sekino, M., Fujiwara, A., Suagya, T., Tanaka, Y., Kumon, K., Agawa, Y., Sagawa, Y., Sano, M., & Sakamoto, T. (2018). Genotyping-by-sequencing for construction of a new genetic linkage map and QTL analysis of growth-related traits in Pacific bluefin tuna. *Aquaculture Research*, 49, 1293–1301. <https://doi.org/10.1111/are.13584>
- Uchino, T., Nakamura, Y., Sekino, M., Kai, W., Fujiwara, A., Yasuike, M., Sugaya, T., Fukuda, H., Sano, M., & Sakamoto, T. (2016). Constructing genetic linkage maps using the whole genome sequence of Pacific bluefin tuna (*Thunnus orientalis*) and comparison of chromosome structure among teleost species. *Advances in Bioscience and Biotechnology*, 7, 85–122. <https://doi.org/10.4236/abb.2016.72010>
- Vasimuddin, M., Sanchit, M., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *IEEE parallel and distributed processing symposium (IPDPS)*. IEEE. <https://doi.org/10.1109/IPDPS.2019.00041>
- Wacker, S., Skaug, H. J., Forseth, T., Solem, Ø., Ulvan, E. M., Fiske, P., & Karlsson, S. (2021). Considering sampling bias in close-kin mark-recapture abundance estimates of Atlantic salmon. *Ecology and Evolution*, 11(9), 3917–3932. <https://doi.org/10.1002/ece3.7279>
- Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, 19(1), 183–194.
- Wang, J. (2022). A joint likelihood estimator of relatedness and allele frequencies from a small sample of individuals. *Methods in Ecology and Evolution*, 13, 2443–2462. <https://doi.org/10.1111/2041-210X.13963>
- Wang, J., & Santure, A. W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, 181, 1579–1594.
- Waples, S. R., & Feutry, P. (2021). Close-kin methods to estimate census size and effective population size. *Fish and Fisheries*, 23(2), 273–293. <https://doi.org/10.1111/faf.12615>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tsukahara, Y., Nakamichi, R., Matsuura, A., Akita, T., Fujiwara, A., & Suzuki, N. (2025). Comparison of kinship-identification methods for robust stock assessment using close-kin mark-recapture data for Pacific bluefin tuna. *Population Ecology*, 1–13. <https://doi.org/10.1002/1438-390X.12205>