SCIENTIFIC COMMITTEE

TWENTY FIRST REGULAR SESSION

Nuku'alofa, Tonga

13 – 21 August 2025

Close-Kin Mark-Recapture of South Pacific Albacore: a feasibility study for application in WCPFC tuna and associated species.

WCPFC-SC21-2025/SA-WP-14

CSIRO[1]and SPC-OFP[2]

[1] The Commonwealth Scientific and Industrial Research Organisation, Australia - Contact: Pierre Feutry (pierre.feutry@csiro.au) or Campbell Davies (campbell.davies@csiro.au)

[2] Oceanic Fisheries Programme (OFP), Pacific Community (SPC), Noumea, New Caledonia – Contact Simon Nicol (simonn@spc.int)

Executive summary

In 2021, the WCPFC Scientific Committee requested an assessment of the feasibility of applying close-kin mark-recapture (CKMR) to South Pacific albacore (SPA) to reduce uncertainties that have been persistent in the past stock assessments for this species (population size, connectivity and appropriate spatial structure). This WP reports upon the completion of two foundational elements of the genetics workflow that are needed for a CKMR application: (1) Quality Control assay; and (2) Kin Identification assay.

Quality Control Assay.

The work focused on evaluating the quality, quantity, and level of cross-contamination of DNA present in samples of SPA collected across the Western and Central Pacific Ocean (WCPO). This served two key purposes: to identify samples most suitable for the resource-intensive kin-finding process and to establish a monitoring system for quality control and assurance of large-scale sampling and data systems required for routine application of CKMR efforts. DNA was extracted from a representative subset of samples collected with a biopsy tool specifically designed for high throughput tissue sampling and extraction— this sample subset covered the major SPA unloading ports and sampling teams and was analysed with a new, low-cost single nucleotide polymorphism (SNP) DarTag assay specifically developed for WCPFC application. Screening of the subset of samples showed that the sampling program implemented across the WCPO performed very well, with less that 4% of the samples failing quality control (QC) checks. This initial result was replicated during the larger-scale kin-finding process, with just over 4% of the 14,763 samples sequenced with the high-resolution SNP assay for kin-finding being discarded due to DNA quality issues.

Kin Identification assay

This work developed a high-resolution DarTag assay designed for kin identification (parent-offspring pairs (POPs), and half-sibling pairs (HSPs)), based on whole genome sequencing of SPA. This new kin-finding SNP assay was tested on ~15, 000 genotyped samples. It performed well, with an estimated loss rate of kin pairs (to avoid false-positive kin) lower than 0.6% when the expected number of false-positives was set to 0.2 pairs. A total of 15 kin pairs were identified, including three POPs and 12 HSPs, from this initial analysis of ~15, 000 samples.

Conclusions

These results clearly demonstrate that two of the major challenges for implementation of a CKMR program to estimate abundance of SPA can be successfully addressed.

- A large-scale tissue sampling and data management program has been successfully established and demonstrated that high quality tissue samples of the thousands of individuals required for CKMR can be obtained, stored and transported for high through-put extraction and genotyping.
- The genotyping, using the new DArTag assay, identifies kin with high confidence and computational efficiency. In particular, the clear separation of HSP from more distant kin demonstrates that the new assay will be sufficiently powerful for SPA and, that the same approach may provide the necessary power for the larger sample sizes required for the tropical tuna, including yellowfin and bigeye.
- CKMR can be feasibly implemented for SPA. Noting that the next stock assessment for SPA is scheduled for 2027, if a CKMR estimator is to be included then completion of sampling and sequencing should be reviewed by SC22 to provide sufficient time for evaluating its readiness for inclusion in the 2027 stock assessment. On the basis of the sequencing and kin results thus far, we do not expect a change to the total number of

samples collected from that recommended in SC20-SA-IP-24 (Tremblay-Boyer-et al 2024).

Recommendations

The SC recommends to WCPFC:

1. Continuation of the sampling program for SPA to facilitate a preliminary absolute abundance estimate for review at SC22.
2. Ongoing application of QC protocols to ensure sample quality is maintained.
3. Completion of sufficient genotyping of samples, kin-finding analyses and associated CKMR modelling to provide an absolute abundance estimate of SPA to be reviewed at SC22.

Introduction

While the application of Close-Kin Mark-Recapture (CKMR) to tuna and associated species in the Western and Central Pacific Fisheries Commission Convention Area (WCPFC-CA) has been considered for some time, its implementation has two significant challenges due to the very large population sizes of tropical tuna species relative to previous applications (e.g., Bravington et al. 2016, Bradford et al. 2018, Hillary et al. 2018, Trenkel et al. 2022): 1) the logistics and costs associated with sampling large numbers (tens or even hundreds of thousands) of tuna across a vast, multi-national geographic area; and 2) the number of pairwise comparisons between individuals in the sample increases quadratically with sample size. To overcome this last challenge, a novel genetic assay with sufficient power is required to find the desired number of kin pairs with high confidence and a very low frequency of false positives.

This paper presents the results of a feasibility study for CKMR in SPA as part of WCPFC SC project 100c and the Climate Science to Ensure Pacific Tuna Access (CSEPTA) project. The study is based on the development of two high-throughput genetic assays, one for assessing DNA quality and cross-contamination, the other being a high-resolution kin-finding SNP assay derived from whole genome sequencing data, which was applied to ~15,000 fish collected in the Western Pacific. The work draws on the information presented in Working Papers SC21-2025/SA-WP-09 and SC21-2025/SA-WP-10, that detail capacity building and sampling activities that underpin this work, and investigations assessing genetic population structure and connectivity of SPA, respectively. These papers in turn build upon work previously presented to the WCPFC SC, including CKMR initial feasibility and design studies for SPA (SC20-2024/SA-WP-09).

Methods

This study consisted of two phases. In the first phase we developed a high-throughput cost-effective genetic assay to assess the quality, quantity and cross-contamination rates of the SPA samples collected to date for CKMR analysis. This served two purposes: (1) to identify samples more likely to be of suitable quality for the cost-intensive kin finding, and (2) provide a routine monitoring assay to track the performance of the sampling program in future CKMR sampling for SPA. This assay was applied to a subset of the available samples, representative of the different sampling locations and samplers involved (Table 1).

The second phase consisted of additional DNA extraction guided by phase one, the development of a high-resolution DarTag assay for kin-finding and the sequencing of ~15,000

with that assay (Table 1). The sample locations used in that second phase are provided in Figure 1. Samples from French Polynesia were not included in this large-scale sequencing as, at the time, they were considered most likely to belong to a different genetic group (Anderson et al. 2019). By not including them, we increased the likelihood of finding kin pairs within the remaining area, which was the primary goal for this feasibility study. Given the results of the population structure and connectivity component of the wider feasibility study (SC21-WP-10), we would now expect to identify kin between French Polynesia, USF and potentially other locations to the west and south. Hence, these samples will be included in the future large-scale genotyping for kin identification.

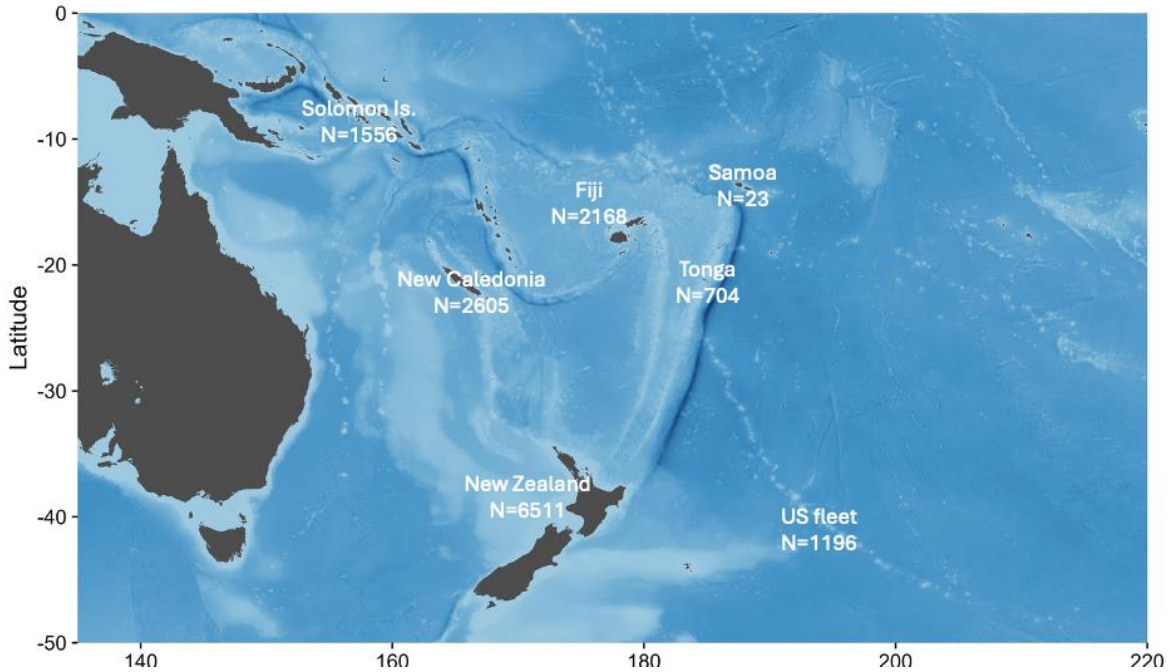

Figure 1. Map of sampling regions.

DNA Extraction

Muscle tissue samples were collected from SPA tuna using single-use biopsy tips developed by CSIRO for gene-tagging of southern bluefin tuna (SBT; Bradford et al., 2016, Preece et al 2016). Immediately after collection, the samples were stored at –20°C and transported on ice to the CSIRO Marine Laboratories in Hobart. In the laboratory, biopsy tips were placed into 96-well deep-well plates, and tissue lysis was carried out directly on the biopsy tips using Proteinase K and Buffer ATL, as provided in the Qiagen QIAamp 96 DNA QIAcube HT Kit (Cat. No. 51331; Qiagen, Hilden, Germany), following a modified version of the kit protocol. The samples were incubated overnight at 56 °C to ensure thorough breakdown of the tissue and efficient release of nucleic acids.

Following tissue digestion, the lysates were centrifuged in their original 96-well plates to pellet any residual tissue debris. This step ensured that debris would not be transferred, thereby reducing the risk of clogging pipette tips or silica membranes during subsequent purification. The cleared lysates were then transferred to a new 96-well plate using the Integra ASSIST PLUS pipetting robot (INTEGRA Biosciences AG, Zizers, Switzerland), which provided consistent sample handling and reduced variability in the pipetting process.

DNA extraction was performed on the Hamilton Microlab STAR automated liquid handling platform (Hamilton Bonaduz AG, Bonaduz, Switzerland), equipped with dual on-deck vacuum

stations. A custom script was developed specifically to enable ultra-high-throughput processing of the modified Qiagen QIAamp 96 DNA QIAcube HT Kit protocol (Cat. No. 51331; Qiagen, Hilden, Germany), allowing extraction of DNA from up to 960 samples/run. Each 96-well plate contained 92 experimental samples with muscle tissue from SPA, 2 control samples from standard reference tissue (SBT), and 2 blank wells serving as negative controls.

DNA Quality Control

Quality control (QC) testing was performed on 25% of the samples from each of the 66 plates available during phase 1 to evaluate DNA extraction efficiency—including DNA concentration, purity, and integrity—and to assess plate-level consistency. QC wells were selected from three representative columns to capture a spatial distribution across each plate that reflects the overall extraction performance. QC sample preparation was automated using the epMotion® 5075 automated liquid handling system (Eppendorf AG, Hamburg, Germany) to ensure reproducibility and reduce manual errors. A total of 1,128 samples selected to represent the range of sampling regions and samplers were sent for sequencing with the high-throughput cost-effective DarTag assay to further assess the suitability of these samples for kin-finding analysis.

For each QC assay, 4µL of DNA was used from a total elution volume of 125µL. DNA concentration and purity was measured using the Multiskan SkyHigh spectrophotometer (Life Technologies Holdings Pte Ltd, Singapore), following a 1:10 dilution of each sample. DNA integrity was evaluated using the Invitrogen E-Gel™ Power Snap Plus Electrophoresis System (Life Technologies Holdings Pte Ltd, Singapore) with precast Invitrogen E-Gel™ 96 Agarose Gels 1% with SYBR™ Safe (Cat No. G720801; Thermo Fisher Scientific, Kiryat Shmona, Israel).

While QC reliably measures extraction efficiency, it also reflects inherent biological variability at the plate level. This subset-based approach assumes that the average tissue quality within each extraction plate is representative, allowing meaningful inference of overall DNA quality across samples on that plate.

During phase 2, another 76 plates were extracted, QC information and sample selection for sequencing with the high-throughput high-resolution DarTag assay for kin finding is provided in table 1.

SNP assays design

The high-throughput cost-effective DarTag SNP assay for DNA QC was designed from the available DArTseq data generated as part of the population structure study (SC21-2025/SA-WP-10). Several hundred SNPs with high minor allele frequency (MAF) were tested with Tag Gen, Diversity Arrays Ltd proprietary automated pipeline for DarTag marker design and 500 loci were retained for design. Synthesis of the oligonucleotides for selected markers was done by Integrated DNA Technologies.

The second DarTag SNP assay was designed to maximise kinship inference power, i.e., selection of loci with high MAF and spread evenly across the genome.

To create a high-resolution SNP assay, additional marker discovery was performed via whole genome sequencing of 188 SPA across the species range on an Illumina Novaseq X platform using pair end run. The average total read count was 109,992,632 and after alignment to in-house reference genome assembly, over 30 million candidate SNPs were identified using FreeBayes (Garrison & Marth, 2012).

Given the high number of good quality markers, we applied stringent parameters to select 5,000 candidates with high MAF and spread uniformly across the genome for marker design, which was done using Tag Gen. Again, the synthesis of the oligonucleotides for selected markers was done by Integrated DNA Technologies.

The synthesised oligonucleotides were pooled with those from the DNA quality and cross-contamination assay, totalling 5,500 loci in the DArTag assay for kin-finding analysis.

Genotyping, data cleaning and kin finding

A total of 1,128 and 14,763 SPA were submitted to DArT for sequencing as part of the first and second phase respectively. The sampling regions are provided in figure 1. The details of the genotyping, data cleaning and kin finding are provided in Appendix 1.

Results

Sample DNA quality

The results of the pre-sequencing DNA quality assessment are presented in Table 1. Notably, French Polynesia exhibited the highest incidence of whole plate flooding, likely due to excessive tissue volumes collected using the biopsy tool. Excessive tissue collection also contributed to variability in DNA purity, as indicated by the large standard deviation, indicating inhibition of DNA purification during the extraction process. New Zealand samples showed the greatest variation in DNA concentration, reflecting inherent biological variability in tissue samples.

Of the initial 1,128 SPA sequenced during phase one, two were removed as duplicates, 15 for high likelihood of being from another species, and 23 likely poor DNA quality, leaving 1088 samples (>96%) for the initial check of quality. Of the initial 525 SNPs in the cleaning check, a total of 116 were removed by allele frequency and Hardy-Weinberg Goodness-of-fit checks, leaving 409.

Of the initial 14,763 samples sequenced during phase two, 130 were removed as duplicates, 271 for high likelihood of being from another species, and 342 for likely poor DNA quality, leaving 13,980 samples for kin-finding. Of the initial 5,109 loci, 1,473 were removed by allele frequency and Hardy-Weinberg Goodness-of-fit checks, leaving 3,636 for kin finding analysis.

Kin-finding

Kin-finding was conducted on the full dataset as part of phase 2. The PLOD_FP values showed three first-order kin were tightly clustered around the expected value for POPs, well below zero, so all three pairs were called as POPs (Fig. 2).

The expected number of false-positives for HSPs was set to 0.2 pairs, which gave an estimated HSP loss-rate of 0.7%. Twelve pairs were above the cutoff and, therefore, called as HSPs (Fig. 2).

Given the low expected false-positive number and low false-negative rate, it is likely that there are no false-positives or false-negatives HSPs in this analysis (Fig. 2).

The samples involved in a kin pair and their associated covariates (length, sampling location, sampling date) are provided in Table 2. The spatial distribution of the kin pair suggests fish movements exist at least between Fiji, New Caledonia, Tonga, New Zealand and Solomon Islands.

Table 1. Summary of DNA extraction results from tissue samples collected during Phase 1 and Phase 2 across seven regions: Fiji, French Polynesia, New Zealand, Solomon Islands, New Caledonia, Tonga, and east of New Zealand by the US fleet. The table reports the total number of samples extracted, the number of flooded and successful extractions, the number of samples quality checked, and those selected for sequencing. Mean DNA concentration (ng/µl) and purity (A260/A280) values are also provided, along with standard deviations to reflect variability. Samples selected for sequencing in Phase 1 were tested for cross contamination between samples to determine the viability of the sampling protocol. The grand total selected for sequencing for CKMR includes individuals extracted during Phase 1 and Phase 2 for the final kin-ship analysis.

Location	Total Extracted	Flooded Extractions	Successful Extractions	Quality Checked	Mean DNA Conc. (ng/µl)	Mean DNA Purity	Total Selected for Sequencing
Phase 1 – sample screening							
Fiji	914	0	914	238	72.5 ± 78.4	1.76 ± 0.73	236
French Polynesia	1196	368	828	144	67.8 ± 47.0	1.93 ± 0.23	118
New Zealand	5828	188	5640	672	46.5 ± 41.3	1.99 ± 0.67	234
Solomon Is.	-	-	-	-	-	-	-
New Caledonia	552	0	552	336	53.2 ± 42.1	2.00 ± 3.68	330
Samoa	-	-	-	-	-	-	-
Tonga	460	0	460	96	82.3 ± 64.4	1.81 ± 0.82	116
US fleet	386	0	386	96	89.6 ± 57.7	1.79 ± 1.28	94
Phase 2 – kin finding							
Fiji	1344	0	1344	308	71.1 ± 69.5	1.88 ± 0.83	2170
French Polynesia	1222	0	1222	307	64.2 ± 64.4	1.85 ± 0.32	0
New Zealand	920	0	920	-	104.0 ± 90.1	1.93 ± 0.20	6511
Solomon Is.	2816	0	2816	404	89.2 ± 62.8	1.94 ± 0.50	1556
New Caledonia	1375	0	1375	357	56.6 ± 47.8	1.93 ± 0.42	2605
Samoa	23	0	23	8	60.6 ± 33.4	1.78 ± 0.13	23
Tonga	344	0	344	24	56.4 ± 44.5	1.76 ± 0.14	704
US fleet	846	0	846	215	95.2 ± 61.3	2.05 ± 1.41	1196

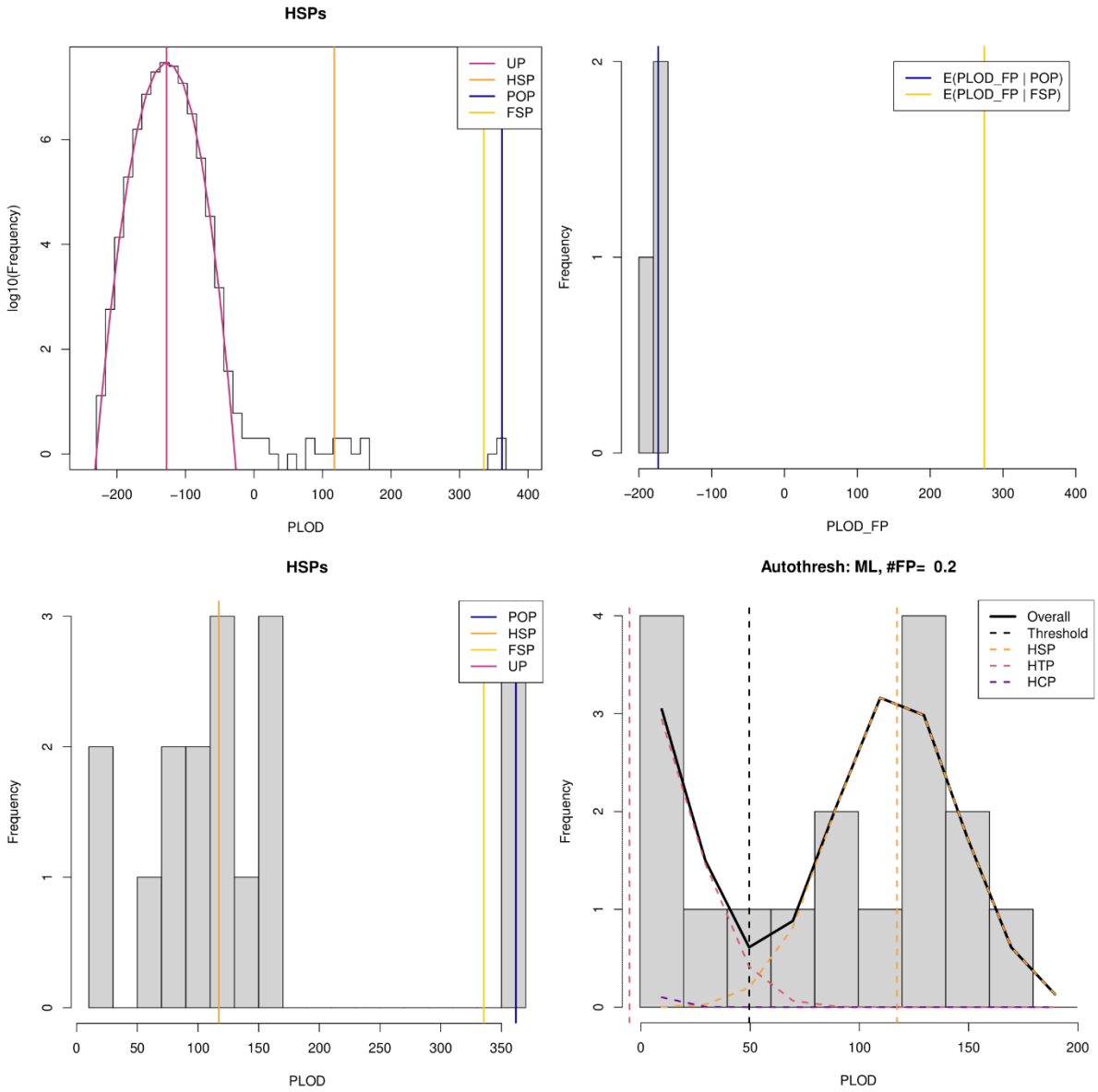

Figure 2. Kinference plots for SPA in the WCPO. Top left: PLOD_HU distribution for all comparisons. Mean and expected distribution is given for UPs, and means are given for HSPs, POPs, and FSPs. Bottom left: PLOD_HU distribution for only those in the 'plausible kin' range. Plausible HSPs are separated from plausible POPs and FSPs by a sizeable gap. Top right: PLOD_FP distribution for all plausible POPs or FSPs. All pairs are tightly clustered around the expected mean for POPs. Bottom right: threshold plot for plausible HSPs. Expected means for HSPs and HTPs are given as coloured dashed lines. The fitted model indicates that only 0.2 third-order or weaker kin are expected to the right of the black dashed line, and only 0.7% of HSPs are expected to fall to the left of the black dashed line. Pairs to the right of the line are taken as HSPs.

Table 2. Kin pairs and associated covariates. "-" indicates no data was available. Length at 50% maturity is ~ 87 cm for SPA (Farley et al, 2014)

Kin pairs ID	Sampling port	Sampling date	Fish length (cm)
POP_01	Solomon Is.	17/10/2024	96
POP_01	Fiji	26/11/2024	102
POP_02	New Zealand	29/02/2024	62
POP_02	Solomon Is.	10/10/2024	91
POP_03	New Zealand	04/03/2023	52
POP_03	New Caledonia	04/10/2024	101
HSP_01	Fiji	16/12/2024	103
HSP_01	New Caledonia	16/02/2025	93
HSP_02	Solomon Is.	21/10/2024	97
HSP_02	New Zealand	07/03/2023	52
HSP_03	New Zealand	19/02/2024	62
HSP_03	Tonga	04/03/2024	95
HSP_04	New Caledonia	27/09/2024	96
HSP_04	Fiji	26/11/2024	90
HSP_05	New Zealand	15/01/2024	62
HSP_05	Tonga	22/07/2024	100
HSP_06	Fiji	27/12/2024	99
HSP_06	Solomon Is.	-	-
HSP_07	New Caledonia	12/10/2024	93
HSP_07	New Zealand	15/03/2024	62
HSP_08	Tonga	04/03/2024	97
HSP_08	New Caledonia	12/02/2025	97
HSP_09	New Zealand	07/03/2023	54
HSP_09	New Zealand	15/01/2024	63
HSP_10	New Zealand	25/03/2024	63
HSP_10	New Zealand	07/03/2023	50
HSP_11	New Zealand	17/02/2025	52
HSP_11	New Zealand	25/03/2024	60
HSP_12	USA	-	-
HSP_12	USA	-	-

Discussion

This study delivered two new DArTag SNP assays. The first allows for cost-effective and rapid monitoring of DNA quality and cross-contamination. The analysis of a subset of the available samples with this assay allowed us to progress to the more expensive kin-finding assay with limited risk of wasting resources. This QC assay can now be used as a routine monitoring assay for the ongoing sampling program in the WCPO. The second DArTag SNP assay provides a cost-effective high-resolution approach to kin-finding. Pairwise comparisons of 13,980 SPA from landings in Fiji, Solomon Islands, New Zealand, New Caledonia, Tonga and the USA, revealed 3 POPs and 12 HSPs. These results will help further refine the implementation of the design study (SC20-2024/SA-IP-24) and sample size requirements from each sampling location for CKMR estimates of abundance and mortality.

Importantly, the resolution of the assay appears sufficient to accommodate at least an order of magnitude more pairwise comparisons, without the need for further assay improvement. This implies that it may be sufficient for populations of the more abundant tropical tunas, such as bigeye and yellowfin tuna.

Conclusions

These results clearly demonstrate that two of the major challenges for implementation of a close-kin program to estimate abundance of SPA can be successfully addressed.

- A large-scale tissue sampling and data management program has been successfully established and demonstrated that high quality tissue samples of the thousands of individuals required for CKMR can be obtained, stored and transported for high through-put extraction and genotyping.
- The genotyping, using the new DArTag assay, identifies kin with high confidence and computational efficiency. In particular, the clear separation of HSP from more distant kin demonstrates that the new assay will be sufficiently powerful for SPA and, that the same approach may provide the necessary power for the larger sample sizes required for the tropical tuna, including yellowfin and bigeye.
- CKMR can be feasibly implemented for SPA. Noting that the next stock assessment for SPA is scheduled for 2027, if a CKMR estimator is to be included then completion of sampling and sequencing should be reviewed by SC22 to provide sufficient time for evaluating its readiness for inclusion in the 2027 stock assessment. On the basis of the sequencing and kin results thus far, we do not expect a change to the total number of samples collected from that recommended in SC20-SA-IP-24 (Tremblay-Boyer-et al 2024).

Recommendations

The SC recommends to WCPFC:

1. continuation of the sampling program for SPA to facilitate a preliminary absolute abundance estimate for review at SC22.
2. ongoing application of QC protocols to ensure sample quality is maintained.
3. completion of sufficient genotyping of samples, kin-finding analyses and associated CKMR modelling to provide an absolute abundance estimate of SPA to be reviewed at SC22.

Acknowledgements

We thank the large number of observers and port samplers who collected samples for this project. We also thank the fishing boat captains and crew as well as the fishing companies that supported our collection efforts at port. Toby Patterson and Ashley Williams provided helpful comments on an earlier draft of this paper.

Funding for this work was provided by the New Zealand Ministry of Foreign Affairs and Trade through the CSEPTA project and CSIRO. We also acknowledge the European Union for their continued support of this work.

References

Anderson, G., Hampton, J., Smith, N., and Rico, C. 2019. Indications of strong adaptive population genetic structure in albacore tuna (*Thunnus alalunga*) in the southwest and central Pacific Ocean. Ecology and Evolution, 9: 10354–10364. 19. https://doi.org/10.1002/ece3.5554

Bradford Russell W., Hill Peta, Davies Campbell, Grewe Peter (2016) A new tool in the toolbox for large-scale, high-throughput fisheries mark-recapture studies using genetic identification. *Marine and Freshwater Research* **67**, 1081-1089. https://doi.org/10.1071/MF14423

Bradford RW, Thomson R, Bravington M, Foote D, Gunasekera R, Bruce BD, Harasti D, Otway N, Feutry P (2018). A close-kin mark-recapture estimate the population size and trend of east coast grey nurse shark. Report to the National Environmental Science Program, Marine Biodiversity Hub. CSIRO Oceans & Atmosphere, Hobart, Tasmania. https://doi.org/10.13140/RG.2.2.26338.58560

Bravington MV, Skaug HJ, Anderson EC (2016). Close-kin mark-recapture. *Statistical Science* 31: 259–275. https://doi.org/10.1214/16-STS552

Farley JH, Hoyle SP, Eveson JP, Williams AJ, Davies CR, Nicol SJ (2014) Maturity ogives for South Pacific albacore tuna (Thunnus alalunga) that account for spatial and seasonal variation in the distributions of mature and immature fish. PLoS ONE 9(1): e83017. https://doi:10.1371/journal.pone.0083017

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*. https://doi.org/10.48550/arXiv.1207.3907

Hillary RM, Bravington MV, Patterson TA, Grewe P, Bradford R, Feutry P, Gunasekera R, Peddemors V, Werry J, Francis MP, Duffy CAJ, Bruce BD (2018). Genetic relatedness reveals total population size of white sharks in eastern Australia and New Zealand. *Scientific Reports* 8: 2661. https://doi.org/10.1038/s41598-018-20593-w

Punt AE, Thomson R, Little LR, Bessell-Browne P, Burch P, Bravington M (2024). Including close-kin mark-recapture data in statistical catch-at-age stock assessments and management strategies. *Fisheries Research* 276: 107057. https://doi.org/10.1016/j.fishres.2024.107057

Tremblay-Boyer L, Bravington MV, Davies C (2024) Updated design models informing the sampling strategy for a Close-kin mark-recapture application to South Pacific albacore. WCPFC-SC20-2024/SA-IP-24 https://meetings.wcpfc.int/node/23132

Trenkel VM, Charrier G, Lorance P, Bravington MV (2022). Close-kin mark–recapture abundance estimation: practical insights and lessons learned. *ICES Journal of Marine Science* 79: 413–442. https://doi.org/10.1093/icesjms/fsac002

Appendix 1

Genotyping

In the first phase, a total of 1,128 SPA were sent to DArT for sequencing with the 500 SNP DarTag assay.

In the second phase, a total of 14,763 SPA from seven sampling regions were sent to Diversity Arrays Ltd for genotyping with the 5,500 DarTag SNP assay. To improve the success rate of this assay, samples were grouped according to DNA quality following agarose gel analysis, so that higher sequencing volume could be applied to lower quality samples. All downstream analyses used DArT 'called' genotypes, with genotypes [0, 1, 2, -] treated as `kinference` 4-way genotypes, [AAO, BBO, AB, OO]. `kinference` 4-way genotypes differ from common biallelic SNPs in that they account for unreadable-but-heritable 'null' ('O') alleles in addition to 'A' and 'B' alleles. If, at one locus, only the A allele is seen for a sample, `kinference` 4-way genotypes treat that sample as either an A allele homozygote ('AA') or an A-null heterozygote ('AO'). This ambiguous genotype is encoded 'AAO'. A similar pattern and encoding holds for the B allele (where 'BB' and 'BO' are encoded 'BBO'). If both the A and B variant are seen at a locus, the sample must have both A and B and cannot have a null allele at that locus (giving the standard 'AB' genotype), and if neither A or B are seen, the sample is taken to be a double-null homozygote ('OO').

Data cleaning

Data cleaning for the first phase dataset consisted of standard pre kin-finding cleaning using kinference (*in prep*). This cleaning consisted of the following nine steps. Note that some steps are repeated: cleaning removes both bad samples and bad loci, but the existence of bad loci can make good samples look bad, and vice-versa. Cleaning therefore occurs in stages, successively removing only the worst samples or loci before arriving at a clean dataset. Population allele frequencies were re-estimated for the remaining samples after every step in which samples were removed. These re-estimates are not listed individually.

1. population 'A', 'B' and 'O' allele frequencies were estimated using a custom maximum-likelihood allele frequency estimator that allows for heritable null alleles

2. loci with an estimated 'O' allele frequency > 0.5 were removed

3. loci with an estimated 'A' allele frequency > 0.95 or < 0.05 were removed

4. loci with an estimated 'B' allele frequency > 0.95 or < 0.05 were removed

5. the most-outlying 2% of loci under a 4-way Hardy-Weinberg Goodness-of-fit test were removed

6. for each sample, a genotype likelihood statistic was calculated to detect samples whose aggregate genotype has a low likelihood of being drawn from a population with allele frequencies equal to those estimated across all samples. Samples with outlying low values in this statistic are often mis-identified samples from another population or species to the overwhelming majority of the other samples, or are cross-contaminated or degraded

7. all remaining loci with P < 0.001 on a 4-way Hardy-Weinberg Goodness-of-fit test (as in step 5) were removed

8. each pair of samples was compared at all loci, and pairs of samples differing at fewer than 200 loci were identified as duplicates. One from each pair of duplicate samples was removed

9. for each sample, a statistic based on the ratio of observed heterozygotes ('AB') to double-nulls ('OO') was calculated. This statistic is designed to detect cross-contaminated samples (with atypically high numbers of heterozygote loci compared to other samples) and degraded samples (with atypically high numbers of null loci compared to other samples). Outlying samples in this statistic were removed

The full dataset during phase 2 was given similar cleaning to the initial dataset. Diversity Arrays Ltd genotypes were interpreted as `kinference` 4-way genotypes, as before. Cleaning processes closely mirrored those for the initial dataset apart from:

5. most-outlying 0.5% of loci (cf. 2% for the initial dataset) were removed based on a 4-way Hardy-Weinberg Goodness-of-fit test.

6. samples with an outlying low value were removed, but the cutoff was very different from that for the initial dataset: the genotype aggregate likelihood statistic is a sum across loci, and the full dataset contains data for an order of magnitude more loci than the initial dataset

10. the genotype aggregate likelihood statistic (as in step 6) was calculated for all samples again, based on the final population allele frequencies for only clean, unduplicated samples. Outlying samples were again trimmed, bringing the final dataset into very close alignment with the *a priori* expected distribution for this statistic.

Of the initial 14896 samples, 130 were removed as duplicates, 261 for outlying `hetzminoo` values, and 476 for outlying `ilglk` values, leaving 13980 samples for kinference. Of the initial 5193 loci, 1541 were removed by allele frequency and Hardy-Weinberg Goodness-of-fit checks, leaving 3652.

Kin-finding

The cleaned dataset was used for kin-finding with package `kinference`. Package `kinference` calculates 'PLOD' statistics, each of which is a likelihood-ratio based statistic for each pair's observed genotypes at each locus under two distinct kinships. The PLOD_HU is statistically optimal for distinguishing half-sibling pairs (HSPs) from unrelated pairs (UPs), and is a useful general-purpose kinship statistic (more closely-related pairs generally score higher on PLOD_HU). The PLOD_FP is statistically optimal for distinguishing parent-offspring pairs (POPs) from full-sibling pairs (FSPs). For each PLOD statistic, the expected mean for specified kin-types can be calculated *a priori*.

First, first-order kin (i.e., POPs or FSPs) were identified visually using the PLOD_HU: POPs and FSPs have *a priori* expected values of PLOD_HU, and samples clustered around these expected values were clearly separate from other pairs. Second, potential POPs and FSPs were distinguished using the PLOD_FP. PLOD_FP is *a priori* expected to be positive for true FSPs and negative for true POPs.

Third, we fit a mixture model to the PLOD_HU for all samples except those already identified as POPs or FSPs. The mixture model models the PLOD_HU density of 2nd-order, 3rd-order, and 4th-order kin (e.g., HSPs, half-thiatic pairs [HTPs], and half-cousin pairs [HCPs]). Given a user-specified expected number of false-positive HSPs, the mixture model provides a PLOD_HU cutoff and an estimated false-negative rate for true HSPs.